BLACKWELL PHILOSOPHY GUIDES

The Blackwell Guide to Ethical Theory

Second Edition

Edited by Hugh LaFollette and Ingmar Persson

WILEY Blackwell

Praise for *The Blackwell Guide to Ethical Theory*, Second Edition

"This outstanding collection contains really valuable entries by a great many leading lights in the field. An extremely useful resource for anyone wanting to acquaint themselves with the central issues in ethical theory today."
—Russ Shafer-Landau, University of Wisconsin-Madison

Praise for the First Edition

"*The Blackwell Guide to Ethical Theory* is, for its size, perhaps the best single-volume resource guide to contemporary ethical theory."
—Robert B. Louden, *APA Newsletter on Teaching Philosophy*, Spring 2001.

"Written by distinguished philosophical experts, these pieces address a full range of issues in ethical and metaethical theory, as well as giving a considered voice to the various defenders of ethical antitheory. As a whole they testify to and exemplify the current vibrancy and richness of work in ethical theory."
—David Archard, Queen's University Belfast

"For students, this Guide provides helpful introductions to themes and topics which they are to study in more detail and lucid surveys of other aspects of ethical theory. It will also be read with profit by academics unfamiliar with the field; some of the papers are not just guides but original contributions to the subject."
—Antony Duff, University of Stirling and University of Minnesota

"The Guide provides a wide-ranging survey of the major topics in contemporary ethical theory and includes a good amount of new and important work. It should be of use to anyone wanting acquaintance with the subject in its current state."
—Barbara Herman, University of California at Los Angeles

Blackwell Philosophy Guides

Series Editor: Steven M. Cahn, City University of New York Graduate School

Written by an international assembly of distinguished philosophers, the *Blackwell Philosophy Guides* create a groundbreaking student resource – a complete critical survey of the central themes and issues of philosophy today. Focusing and advancing key arguments throughout, each essay incorporates essential background material serving to clarify the history and logic of the relevant topic. Accordingly, these volumes will be a valuable resource for a broad range of students and readers, including professional philosophers.

1 The Blackwell Guide to Epistemology
edited by John Greco and Ernest Sosa

2 The Blackwell Guide to Ethical Theory, Second Edition
edited by Hugh LaFollette and Ingmar Persson

3 The Blackwell Guide to the Modern Philosophers
edited by Steven M. Emmanuel

4 The Blackwell Guide to Philosophical Logic
edited by Lou Goble

5 The Blackwell Guide to Social and Political Philosophy
edited by Robert L. Simon

6 The Blackwell Guide to Business Ethics
edited by Norman E. Bowie

7 The Blackwell Guide to the Philosophy of Science
edited by Peter Machamer and Michael Silberstein

8 The Blackwell Guide to Metaphysics
edited by Richard M. Gale

9 The Blackwell Guide to the Philosophy of Education
edited by Nigel Blake, Paul Smeyers, Richard Smith, and Paul Standish

10 The Blackwell Guide to Philosophy of Mind
edited by Stephen P. Stich and Ted A. Warfield

11 The Blackwell Guide to the Philosophy of the Social Sciences
edited by Stephen P. Turner and Paul A. Roth

12 The Blackwell Guide to Continental Philosophy
edited by Robert C. Solomon and David Sherman

13 The Blackwell Guide to Ancient Philosophy
edited by Christopher Shields

14 The Blackwell Guide to the Philosophy of Computing and Information
edited by Luciano Floridi

15 The Blackwell Guide to Aesthetics
edited by Peter Kivy

16 The Blackwell Guide to American Philosophy
edited by Armen T. Marsoobian and John Ryder

17 The Blackwell Guide to Philosophy of Religion
edited by William E. Mann

18 The Blackwell Guide to the Philosophy of Law and Legal Theory
edited by Martin P. Golding and William A. Edmundson

19 The Blackwell Guide to the Philosophy of Language
edited by Michael Devitt and Richard Hanley

20 The Blackwell Guide to Feminist Philosophy
edited by Linda Martín Alcoff and Eva Feder Kittay

21 The Blackwell Guide to Medical Ethics
edited by Rosamond Rhodes, Anita Silvers, and Leslie P. Francis

The Blackwell Guide to Ethical Theory

Second Edition

Edited by
Hugh LaFollette and Ingmar Persson

WILEY Blackwell

This edition first published 2013

Edition history: Blackwell Publishing Ltd (1e, 2000).

Blackwell Publishing was acquired by John Wiley & Sons in February 2007. Blackwell's publishing program has been merged with Wiley's global Scientific, Technical, and Medical business to form Wiley-Blackwell.

Registered Office
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

Editorial Offices
350 Main Street, Malden, MA 02148-5020, USA
9600 Garsington Road, Oxford, OX4 2DQ, UK
The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

For details of our global editorial offices, for customer services, and for information about how to apply for permission to reuse the copyright material in this book please see our website at www.wiley.com/wiley-blackwell.

Library of Congress Cataloging-in-Publication Data
The Blackwell guide to ethical theory. – Second edition / edited by Hugh LaFollette, Ingmar Persson.
pages cm. – (Blackwell philosophy guides ; 2)
Includes bibliographical references and index.
ISBN 978-1-4443-3009-0 (pbk. : alk. paper) 1. Ethics. I. LaFollette, Hugh, 1948– II. Persson, Ingmar.
BJ1012.B536 2013
171–dc23
2012042778

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

Cover image: Caspar David Friedrich, *Evening*, c.1820. State Museum, Hanover / akg-images
Cover design by Nicki Averill

Set in 10/13 pt Galliard by Toppan Best-set Premedia Limited

1 2013

Contents

Notes on Editors and Contributors

Editors

Hugh LaFollette is Marie E. and E. Leslie Cole Chair in Ethics at the University of South Florida St. Petersburg. He was editor-in-chief for the *International Encyclopedia of Ethics*, a nine-volume work recently published by Wiley-Blackwell. He is author of three books and numerous essays in ethics and political philosophy. hughlafollette@tampabay.rr.com

Ingmar Persson is Professor of Practical Philosophy, University of Gothenburg, and Distinguished Research Fellow, Oxford Uehiro Centre for Practical Ethics. His publications include *The Retreat of Reason: A Dilemma in the Philosophy of Life*, *Unfit for the Future: The Need for Moral Enhancement* (coauthored with Julian Savulescu), and *From Morality to the End of Reason: An Essay on Rights, Reasons and Responsibility* (forthcoming). ingmar.persson@phil.gu.se or ingmar.persson@philosophy.ox.ac.uk

Contributors

Simon Blackburn is a fellow of Trinity College, Cambridge, Distinguished Research Professor at the University of North Carolina, Chapel Hill, and Professor at the New College of the Humanities. He was a fellow of Pembroke College Oxford from 1970 to 1990, and is a fellow of the British Academy and Honorary Foreign Member of the American Academy of Arts and Sciences. His books include *Spreading the Word*, *Ruling Passions*, *Think*, *Truth: A Guide for the Perplexed*, *How to Read Hume* and *Plato's Republic*. Collections of papers are *Essays in Quasi-Realism* and *Practical Tortoise Raising*. swb24@cam.ac.uk

John M. Doris is Professor in the Philosophy–Neuroscience–Psychology Program, Washington University in St. Louis. He authored *Lack of Character* and has authored or coauthored papers for numerous journals and books. Doris has held fellowships from Michigan's Institute for the Humanities, Princeton's University Center for Human Values, the National Humanities Center, the American Council of Learned Societies, and (three times) the National Endowment for the Humanities. He has been awarded the Society for Philosophy and Psychology's Stanton Prize for interdisciplinary research in philosophy and psychology. jdoris@artsci.wustl.edu

R.G. Frey was Professor of Philosophy at Bowling Green State University. He was the author of numerous articles and books in ethical theory, applied ethics, the history of ethics, and social/political theory. New works of his awaiting publication are books on Joseph Butler, an edition of Butler's ethical writings, a volume of essays on topics in applied ethics, and a volume on utilitarianism. His *Euthanasia and Physician-Assisted Suicide*, with Gerald Dworkin and Sissela Bok, was published in 1998.

Thomas E. Hill Jr, currently Kenan Professor of Philosophy at the University of North Carolina at Chapel Hill, previously taught at UCLA. His essays on moral and political philosophy are collected in *Autonomy and Self-Respect*, *Dignity and Practical Reason in Kant's Moral Theory*, *Respect, Pluralism, and Justice: Kantian Perspectives*, *Human Welfare and Moral Worth: Kantian Perspectives*, and *Virtue Rules and Justice: Kantian Aspirations.* He coedited (with Arnulf Zweig) *Kant's Groundwork for the Metaphysics of Morals* and edited *The Blackwell Guide to Kant's Ethics.* thill@email.unc.edu

Brad Hooker is a professor in the philosophy department at the University of Reading. He has published widely on moral philosophy. His most significant publication is *Ideal Code, Real World: A Rule-Consequentialist Theory of Morality.* b.w.hooker@reading.ac.uk

Alison M. Jaggar is A&S College Professor of Distinction in Philosophy and Women and Gender Studies at the University of Colorado at Boulder and Research Coordinator at the University of Oslo's Centre for the Study of Mind in Nature. Jaggar is author and editor of numerous articles and books. Currently, she is working with a multidisciplinary international team seeking to produce a new poverty metric capable of revealing the gendered dimensions of global poverty. In addition, Jaggar is exploring the potential of a naturalized approach to moral epistemology for addressing moral disputes in contexts of inequality and cultural difference. alison.jaggar@colorado.edu

Richard Joyce has held positions in the United Kingdom and Australia, and is currently Professor of Philosophy at Victoria University of Wellington (New

Zealand). He is author of *The Myth of Morality*, *The Evolution of Morality*, and numerous papers on metaethics and moral psychology. He is coeditor (with Simon Kirchin) of *A World Without Values: Essays on John Mackie's Moral Error Theory*. richard.joyce@vuw.ac.nz

F.M. Kamm is Lucius Littauer Professor of Philosophy and Public Policy, Harvard Kennedy School, and Professor of Philosophy, Harvard University. She has authored *Creation and Abortion*; *Morality, Mortality*, vols. 1 and 2; *Intricate Ethics*; *Ethics for Enemies: Terror, Torture and War*; *The Moral Target: Aiming at Right Conduct in War and Other Conflicts*; and numerous articles. She serves on the editorial board of *Philosophy & Public Affairs* and on the Advisory Committee of the Edmond J. Safra Ethics Center. She has held Guggenheim and NEH Fellowships and is a fellow of the American Academy of Arts and Sciences. frances_kamm@harvard.edu

Ron Mallon is an associate professor of philosophy and director of the Philosophy-Neuroscience-Psychology Program at Washington University in St. Louis. He has codirected two NEH Summer Institutes on Experimental Philosophy; he has been the recipient of an American Council of Learned Societies Fellowship, a Laurence S. Rockefeller Visiting Fellowship at Princeton's University Center for Human Values, and a Charlotte W. Newcombe Doctoral Dissertation Fellowship. His research explores the intersection of culture and the mind in the philosophy of psychology, experimental philosophy, and moral psychology. rmallon@wustl.edu

Jeff McMahan is Professor of Philosophy at Rutgers University. He is the author of *The Ethics of Killing: Problems at the Margins of Life* and *Killing in War*. A collection of his essays, *The Values of Lives*, is also forthcoming. mcmahan@philosophy.rutgers.edu

David McNaughton is Professor of Philosophy at Florida State University, and Professor Emeritus at Keele University. He is the author of *Moral Vision* and (with Eve Garrard) of *Forgiveness*, and of a number of papers on ethics, philosophy of religion, and the relations between the two. He is currently writing a book with Piers Rawling on their approach to practical reasons. dmcnaughton@fsu.edu

Jan Narveson, BA (Chicago), PhD (Harvard) (and FRSC, and OC) is Distinguished Professor Emeritus of Philosophy at the University of Waterloo in Ontario, Canada. He is the author of over two hundred papers in philosophical periodicals and anthologies, and of several books: *Morality and Utility*; *The Libertarian Idea*; *Moral Matters*; *Respecting Persons in Theory and Practice*; and *You and The State*; also, with Marilyn Friedman, *Political Correctness.* jnarveso@uwaterloo.ca

Derek Parfit was born in 1942. His main areas of interest are ethics, metaethics, normativity, and metaphysics. His first book, *Reasons and Persons*, was published

in 1984, and the two volumes of his second book, *On What Matters*, were published in 2011. Parfit is an emeritus fellow at All Souls College, Oxford, and a regular visiting professor to the Departments of Philosophy of Harvard, Rutgers, and New York University. derek.parfit@all-souls.ox.ac.uk

Philip L. Quinn was, for many years, the John A. O'Brien Professor of Philosophy at the University of Notre Dame. He is the author of *Divine Commands and Moral Requirements*, as well as numerous papers in the philosophy of religion, philosophy of science, ethics, metaphysics, and the history of philosophy. He was coeditor of *A Blackwell Companion to the Philosophy of Religion*.

Piers Rawling is Professor and Chair of Philosophy at Florida State University. He has wide-ranging interests, and has published papers on decision theory, ethics (with David McNaughton), metaphysics, philosophy of action, language, mind, science, and quantum computing (with Stephen Selesnick). He is coeditor (with Alfred Mele) of *The Oxford Handbook of Rationality*. prowling@fsu.edu

Ingrid Robeyns is Professor of Practical Philosophy at the Erasmus University, Rotterdam. Her main areas of research are theories of justice, the capability approach, and normative political philosophy related to economic questions and to the family. robeyns@fwb.eur.nl

Geoffrey Sayre-McCord is the Morehead Alumni Distinguished Professor at the University of North Carolina at Chapel Hill. In addition to working in metaethics, he writes on moral theory, epistemology, and the history of philosophy. sayre-mccord@unc.edu

William R. Schroeder is Emeritus Professor at the University of Illinois at Urbana-Champaign. In addition to various articles on Nietzsche, Sartre, interpretation theory, and ethics, he has written *Sartre and His Predecessors: The Self and The Other*, about intersubjectivity; and *Continental Philosophy: A Critical Approach*, a full survey of the field. He also coedited *A Blackwell Companion to Continental Philosophy*. Currently he is working on books on Nietzsche's ethics, on Scheler's ethics, and on editing the second edition of *A Blackwell Companion to Continental Philosophy*. wschroed@illinois.edu

Michael Slote is UST Professor of Ethics at the University of Miami. A member of the Royal Irish Academy and former Tanner lecturer, his most recent books include *The Ethics of Care and Empathy*; *Moral Sentimentalism*; and *Education and Human Values*. His newest book (forthcoming) is *From Enlightenment to Receptivity: Rethinking Our Values*. mslote@miami.edu

Michael Smith is McCosh Professor of Philosophy at Princeton University. He is the author of *The Moral Problem*, *Ethics and the A Priori: Selected Essays*

on Moral Psychology and Meta-ethics, and the coauthor, with Frank Jackson and Philip Pettit, of *Mind, Morality, and Explanation: Selected Collaborations*. msmith@princeton.edu

Elliott Sober is Hans Reichenbach Professor of Philosophy and William F. Vilas Research Professor at University of Wisconsin-Madison. His research is in the philosophy of science, especially in the philosophy of evolutionary biology. Sober's books include *The Nature of Selection – Evolutionary Theory in Philosophical Focus*, *Reconstructing the Past – Parsimony, Evolution, and Inference*, *Philosophy of Biology*, *Unto Others – The Evolution and Psychology of Unselfish Behavior* (with David Sloan Wilson), *Evidence and Evolution – The Logic Behind the Science* (2008), and most recently, *Did Darwin Write the Origin Backwards?*. ersober@wisc.edu

L.W. Sumner is University Professor Emeritus in the Department of Philosophy at the University of Toronto. He is the author of five books: *Abortion and Moral Theory*; *The Moral Foundation of Rights*; *Welfare, Ethics, and Happiness*; *The Hateful and the Obscene: Studies in the Limits of Free Expression*; and most recently, *Assisted Death: A Study in Ethics and Law*. He is a fellow of the Royal Society of Canada and recipient of the 2009 Molson Prize in Social Sciences and Humanities from the Canada Council for the Arts. sumner@chass.utoronto.ca

Introduction

Hugh LaFollette and Ingmar Persson

Contemporary moral philosophers entertain theories about human nature, explore the nature of value, discuss competing accounts of the best ways to live, ponder the connections between ethics and human psychology, and discuss practical ethical quandaries. Broadly conceived, these are the same issues ancient philosophers discussed. However, the precise questions contemporary philosophers ask, the distinctions we make, the methods we employ, and the knowledge of the world and of human psychology we use in framing and evaluating ethical theories, often only faintly resemble those of our philosophical predecessors.

Nonetheless, current ethical theories are shaped by our predecessors. We wrestle with the questions they posed. We ask the questions we ask, in the ways we ask them, because of their philosophical successes and failures. Their debates were likewise shaped by the questions posed by their predecessors. The connection between our, their, and their predecessors' questions explains why we have a history of ethics, why we all participate in the same debate. The differences between their debates and ours reflect the ways ethics has evolved. This is as it should be: the debates are similar because we are all looking for better ways to relate to each other; different because with time and the benefits of hindsight, we should better understand ourselves, our place in the world, and our relationships to others.

We can divide their questions and ours into three broad categories: metaethics, normative ethics, and practical ethics. Here are examples of each.

Metaethics: What is the status of moral judgments? Are they statements of fact or expressions of attitudes? If they are statements of fact, are these facts subjective or objective? Are they statements about a normative or evaluative realm

The Blackwell Guide to Ethical Theory, Second Edition. Edited by Hugh LaFollette and Ingmar Persson.

distinct in kind from the natural world? Is moral agreement possible or do we land in relativism?

Normative ethics: What is the best way, broadly understood, to live? Are there general principles, rules, guidelines that we should follow, or virtues that we should inculcate, that help us distinguish right from wrong and good from bad?

Practical (or applied) ethics: How should we behave in particular situations: when should we tell the truth, under what circumstances can or should we go to war, what is the best way to organize society, how should we relate to the environment and to animals?

The first part of this book contains essays in metaethics and moral epistemology: they discuss the nature and status of ethics and our knowledge of moral matters. The third and largest part contains essays in normative ethics: they offer competing accounts of how we should live. In-between them there is a part with essays on factual matters of relevance to ethics, such as how the psychology of beings must be if they are to be capable of developing and following ethical norms. It must arguably be such that they are capable of being responsible for their actions and being altruistically motivated. This book does not cover practical ethics.

The idea that ethics, like Gaul, is divided into three parts reveals the ways in which ethics as a discipline has evolved. For this is not a distinction the ancients made. Likely they would have seen it as a contrivance, carving nonexistent joints in the moral universe. Still, we can, without undue violence to their views, classify their discussions into these three camps. Plato's theory of the forms (as traditionally understood) could be seen as the first attempt at defending moral realism and offering an objective ground for moral truths. Aristotle's account of the virtues is an early example of virtue theory. And Plato's proposed structure of the state could be envisaged as an early exercise in practical ethics. As long as we use these distinctions as simply a convenient way of distinguishing the kinds of questions they asked, then likely they would not take umbrage.

In the middle of the twentieth century philosophers, however, did not see these distinctions as merely part of a useful classificatory scheme, but as separating ethics into three wholly distinct disciplines, where metaethics was the primary discipline and the only one which constituted "real" philosophy. Philosophy in the English-speaking countries had taken a decidedly "linguistic turn," and the analysis of language, in particular everyday language, was seen as the chief occupation of a philosopher. The British philosopher A.J. Ayer boldly proclaimed that "the propositions of philosophy are not factual, but linguistic . . ." (1946/1952: 57). This was a view of philosophy which originated in the Vienna Circle in the 1930s.

Accordingly, moral philosophers were almost exclusively concerned with the analysis of ethical terms. Typical titles of the period are Charles Stevenson's *Ethics and Language* (1944) and R.M. Hare's *The Language of Morals* (1952). Moral philosophers at the time thought they could fruitfully engage in metaethics without the slightest interest in or acquaintance with normative or practical ethics. Indeed, most would not even consider normative ethics as part of ethics. Not surprisingly,

virtually no one then would have envisioned practical ethics as we now know it. Still less would they be tempted to think of it as philosophy.

Present-day moral philosophers engage in activities that fall into all three categories. Unlike their mid-century predecessors, they reject the idea that moral philosophy is equivalent to metaethics. They also recognize the great relevance that some empirical matters have for ethics, as will be evident from the essays in Part II. Moreover, they are disinclined to think that the three ethical categories make up three wholly independent inquiries. For instance, Stephen Darwall rejects any clear separation between metaethics and normative ethics (1998: 12), while Shelly Kagan not only eschews the distinction between metaethics and normative ethics (1998: 7), he also renounces any firm distinction between normative and practical ethics (1998: 5). In addition, as you will notice when reading the essays, distinctions between the types of normative theory are likewise blurry and should not be taken as indicating more than a vague family resemblance.

The Essays

Metaethics and Moral Epistemology

Part I begins with two essays discussing the meaning of moral judgments: Are they true or false descriptions of states of affairs or expressions of noncognitive attitudes like desires and emotions? If they describe states of affairs, are these objective, or do they have to do with subjective attitudes? If they describe or express attitudes, will not moral judgments be relative to individuals or cultures?

Michael Smith takes "Moral Realism" to be the doctrine that (1) moral judgments are capable of being true and false, and that (2) some of those judgments are, in fact, true. Realism is best understood in contrast with two alternatives: nihilism and expressivism. Both alternatives agree that moral claims cannot be true: expressivists by denying (1) and nihilists by denying (2).

This way of characterizing realism is, however, problematic because there is a popular "minimalistic" conception of truth, according to which saying that a sentence is true is just expressing agreement with it. In this sense, expressivists can claim that moral claims are capable of being true. A better way of characterizing expressivism is by saying that it endorses *internalism*, the doctrine that there is a necessary connection between making a moral judgment and being motivated to act in accordance with it, for instance, between judging that it is wrong to torture babies and being averse to doing it. Although Smith rejects expressivism, he thinks expressivists are right about internalism. He accepts an internalist naturalistic moral realism which claims that judging an act to be right is judging that it is such that in conditions of ideal reflection it induces a desire that it be performed. Internalist naturalistic realism stands in opposition not only to externalist naturalistic realism, but also to nonnaturalistic realism, which takes moral judgments to refer

to properties that are distinct in kind and irreducible to the empirical properties investigated by science.

But will the desires of all of us converge under the conditions of ideal reflection? Smith thinks that a presumption to this effect is part of the meaning of our moral terms. So, he is to that extent a nonrelativist, but he admits that this presumption might turn out to be false. If so, relativism or nihilism will have the final word.

In "Relativism" Simon Blackburn agrees with Smith that, because of the psychological and cultural differences between human beings, relativism poses a threat to moral realism. The remedy Blackburn proposes is, however, to give up the realist idea that moral judgments purport to represent how the world is and an accompanying substantive conception of truth, according to which they are true when they succeed in so doing. Instead we should accept the expressivist claim that moral judgments express the speakers' desires and emotions and a minimalist conception of truth, according to which saying that moral judgments are true is merely to endorse the attitudes they express. Contrary to what many think, Blackburn believes that expressivism can give satisfactory sense to the idea that moral norms can be objective: they are objective if they are not biased or blind to relevant facts. They can also possess an appropriate authority if we regard it as not optional or permissible to embrace conflicting norms.

In "Moral Agreement" Derek Parfit also grapples with the question of whether in ideal conditions – in which we are all adequately informed about relevant empirical facts, thinking clearly, and not subject to any distorting emotions – our moral beliefs will tend to converge. Parfit accepts a form of nonnaturalistic moral realism, according to which we have intuitive knowledge of the basic norms of rationality and morality. (McNaughton and Rawling defend a similar kind of view in "Intuitionism.") This view would be dubious if even under ideal conditions there were deep and widespread divergences in respect of the norms that we endorse. But Parfit contends that this does not seem to be so. In many cases moral disagreements can be put down to differences as regards nonmoral beliefs, for example about matters of religion. In others they are due to distorting factors, such as when the selfishness of the rich makes them underestimate the extent to which they ought to aid the poor. But here, Parfit claims, people also disagree because they think that moral norms are more precise than they actually are. It is unlikely that there is a precise amount that the rich are required to give; in many cases it is indeterminate whether or not the rich have given as much as is required, just as in many cases it is indeterminate whether or not a man is bald. Parfit is thus hopeful that when we get a clearer picture of the reasons for moral disagreement, we could reasonably conjecture that the convergence of our moral belief in ideal conditions would be so substantial that his nonnaturalistic moral realism is not jeopardized.

In "Divine Command Theory" Philip L. Quinn offers another account of the grounds of the objectivity of morality. He defends the claim that morality depends on God. His aim is not to convince atheists of the truth of his view – that would depend on him first convincing them of the existence and nature of God. His aim

is more modest: to show that such beliefs constitute a defensible theory of ethics. Quinn offers and refines a version of the Divine Command Theory, and then responds to assorted criticisms which purportedly show that the very idea of a divine command ethic is indefensible. He ends by offering what he sees as a "cumulative case argument" for the suggestion that categorical moral requirements cannot exist unless there is a deity that intends them.

By a "moral intuition" Jeff McMahan means a moral belief or judgment that is not the result of an inference. He uses the term in a broader, metaethically neutral way rather than in any of the more specific ways it has been used by non-naturalistic moral realists like Parfit, and McNaughton and Rawling. In his essay "Moral Intuition" McMahan sketches a *foundationalist* moral epistemology which takes as its starting point our intuitions about particular moral matters – for example, that it is wrong to kill innocent human beings – and works its way bottom-up to more general moral principles. The principles extracted then need to be confirmed by a top-down procedure of checking how they square with other particular intuitions. This methodology may seem indistinguishable from what John Rawls famously called the search for a *reflective equilibrium*. But Rawls's procedure has usually been interpreted as a coherentist approach, whereas McMahan takes his proposal to be foundationalist. This is because McMahan believes that we have some moral intuitions that we regard as so certain that we would not be willing to surrender them to achieve greater coherence. What ultimately provides the justificatory foundation in his view are the general principles we abstract from our more particular intuitions, but we must have recourse to these intuitions in order to discover the principles: "The order of discovery is the reverse of the order of justification," as he puts it.

Factual Background of Ethics

Part II opens with an essay by Richard Joyce in which he approaches "Ethics and Evolution" from both an empirical and a philosophical perspective. The main issue of the empirical approach is the truth of *moral nativism*, the doctrine that moral judgments are not cultural products but products of a biological adaptation: an innate trait favored by evolution because it provides humans with a reproductive advantage. One problem confronting this hypothesis concerns what precisely "moral judgments" are. Joyce tentatively suggests that they are judgments that involve a notion of justification and that possess a "categorical" authority which gives them an independence of subjective attitudes. Another major problem concerns lack of evidence for this hypothesis that rules out competing hypotheses, such as the hypothesis that the disposition to make moral judgments is a by-product of other biological adaptations.

However, assuming that moral nativism has enough going for it, the philosophical question arises as to what implications it has for ethics. The implications can concern metaethics, normative ethics, or practical ethics, and they can be either

vindicating or debunking. Joyce ends by outlining some such implications and the objections they face.

In the second essay of this part Elliott Sober discusses a long-standing worry about "Psychological Egoism." Philosophers since Socrates have worried that all human motivation is ultimately self-interested. If this is the case, it seems that genuinely moral behavior is impossible because it presupposes that we are capable of being motivated by altruism, that is, by a concern about the well-being of others *for its own sake*. Sober argues that the philosophical arguments against psychological egoism, like the famous argument by Joseph Butler, fail. Nor do everyday or scientific observations settle the issue. However, Sober believes that an argument from evolutionary biology tells against psychological egoism, albeit not decisively. Humans can grow, learn, and flourish only if they are given suitable parental care and guidance. But a pure egoistic hedonist will be a less reliable parent than either a purely altruistic parent or a parent acting from a plurality of motives that includes altruism. This is so because altruistic parents are capable of being motivated to help their offspring *directly* by a belief that it needs help, whereas in order for hedonist parents to be motivated to help they must have an additional belief that the help would provide the parents themselves with some benefit.

It is customary in recent empirical research on moral judgment to distinguish between, on the one hand, "system 1" processes that are phylogenetically ancient, fast, unconscious, emotional, intuitive, and effortless and, on the other hand, "system 2" processes that are phylogenetically recent, slow, conscious, voluntary, and effortful. The claim of what Ron Mallon and John Doris in "The Science of Ethics" call *psychological intuitionism* is that system 1 processes exercise a dominating influence on the formation of moral judgment. This has nurtured skepticism about its rationality. Mallon and Doris critically review some arguments in favor of the dominance of system 1, such as the argument that this dominance is necessitated by our limited cognitive resources and the appeal to rationalizing confabulation that reasoning has been found to engage in. They suggest that psychological intuitionism makes the mistake of overlooking or downplaying the role of individual experience and transmitted culture in the shaping of our moral thinking.

It is a well-known claim that morality presupposes responsibility; that it would not make sense to hold there to be some acts that people morally ought to perform unless they could be responsible for their acts. It is also a well-known claim that people cannot be responsible because this is incompatible with determinism, the doctrine that everything that happens has a sufficient cause. In "The Relevance of Responsibility to Morality" Ingmar Persson rejects this incompatibility. He argues that what is necessary for responsibility is that we can conduct deliberation on the basis of reasons for action and that this only requires that we cannot reliably *predict* the outcomes of our deliberations because we are necessarily unaware of some of their causes, even if there are such causes. Apart from its relevance for deliberation, Persson claims that assumptions about responsibility are relevant for some of the *content* of commonsense morality: for deontological doctrines, such as the act–omission doctrine, and for ascriptions of desert. He argues that in both cases

mistaken assumptions about responsibility are involved; assumptions to the effect that responsibility is based upon causation and that it can be ultimate, respectively. His overall suggestion is that an exploration of responsibility can necessitate a revision of the content of morality, but it will not undercut all moral norms.

Normative Ethics

Large portions of the history of normative ethics have been dominated by two traditions: consequentialism and deontology. Accordingly, Part III starts with essays on approaches that firmly belong to one of these two traditions, two essays on consequentialism and three essays on deontology. However, in the course of time the line between these two traditions has become increasingly blurry and, for instance, contractarianism could incorporate features from either of them.

The most dominant forms of consequentialism have been utilitarian. Utilitarianism has played a pivotal role in the history of ethics. Not only has the theory enjoyed considerable support among philosophers, it has also played a central role as a foil for other theories. Many deontological theories were developed and refined through their attempts to distinguish themselves from utilitarianism. Historically the most widely advocated form of utilitarianism was "Act-Utilitarianism." R.G. Frey explains that the original appeal of act-utilitarianism was the belief that the theory offers a (1) relatively simple and (2) easily applied moral theory and, hence, could be used by most people to make everyday moral decisions.

Partly because of its once dominant role, act-utilitarianism was subject to fierce criticism by its detractors, and subsequent scrutiny and revision by its defenders. It became clear to both parties that the theory was neither simple nor easily applied. In its aim to avoid critics and satisfy adherents, the theory has undergone substantial change. One of most interesting developments was R.M. Hare's indirect utilitarianism, which distinguished between judgments at the critical and at the intuitive level. Utilitarianism is supposed to govern judgment at the critical level, but not at the intuitive (everyday) level. These modifications were supposed to bring act-utilitarianism closer to commonsense morality and, therefore, make it more defensible. Although this move is well motivated, in the end, Frey argues, it does not work. Nonetheless, it points in the right direction: utilitarians must distinguish between the theory as an account of the right and as a decision procedure. Act-utilitarianism, properly understood, is simply an account of right action, not a decision procedure. It recommends the adoption of such decision procedures and the development of such character traits that makes us most likely to act in ways that promote the greatest utility.

The second prominent form of consequentialism is "Rule-Consequentialism," which claims that actions are right, not because they have the best consequences, but because they spring from a set of rules that have the best consequences. Brad Hooker claims this sort of theory is more plausible than act-utilitarianism because it does not require us to break moral rules for the sake of only marginally better

consequences; nor does it demand exceedingly much in the way of self-sacrifice to aid others. Also, by taking the goodness of consequences to depend not merely upon how much utility is produced, but also upon how fairly it is distributed, Hooker avoids some of the objections that have traditionally been leveled at utilitarianism.

According to Hooker, rule-consequentialism is better understood as the doctrine that an act is right if it is in accordance with rules the *acceptance or internalization* of which has best consequences, rather than as the doctrine that it is right if it is in accordance with rules *compliance* with which has the best consequences. This enables him to answer the well-known objection that rule-consequentialism collapses into act-consequentialism. Hooker goes on to tackle the tricky problem of deciding how widely accepted the rule-consequentialist should assume the rules to be. We cannot realistically expect *everyone* to internalize the principles, but the rate of internalization has to be very high to justify the idea that the principles should hold good for the *whole* of society.

The contrast between consequentialism and deontology comes out starkly in F.M. Kamm's essay "Nonconsequentialism." She characterizes nonconsequentialism as (1) setting *constraints* on what we can do in our quest to pursue either the impersonal good or our own good – for example, we are not permitted to harm nonconsenting people as means to these ends – and as (2) granting *prerogatives* for each individual to set aside the goal of maximizing the good when this requires extensive sacrifices.

Nonconsequentialists differ on the question of whether constraints are to be regarded as absolute or as having thresholds to the effect that if the good produced is very much greater than the harm inflicted, it is permissible to inflict the harm – for example, to kill one to save thousands. Kamm claims, however, that *no* amount of *smaller* harms, such as sore throats, can be as morally important as a great harm, like death.

Another matter of controversy among nonconsequentialists is whether we are permitted to violate constraints in order to prevent more constraints being violated – for example, kill one to prevent five from being killed. Kamm's answer is that we have an *inviolability* which makes this impermissible – unless the number of deaths is above the threshold. The thresholds of constraints in conjunction with prerogatives give rise to a nontransitivity of permissibility which Kamm also tries to explain: it might be permissible to violate a constraint to promote a great good, and permissible to set aside this great good because it involves too great a personal cost, yet not permissible to violate the constraint to save oneself this cost.

Appeals to intuitions play an important role in Kamm's account of morality, as they do in many accounts. But "Intuitionism" usually designates a special form of nonnaturalist realism that was popular in England between the wars, but then fell into disrepute. Yet, in the last couple of decades it has made a theoretical comeback. Intuitionism always had one enormous asset: it appeared to accommodate ordinary moral thinking. Despite this asset, many philosophers thought (and still think) the theory plagued with insurmountable difficulties; most especially, that it

is burdened by belief in mysterious nonnaturalistic moral properties, is nonexplanatory, and wholly unsystematic. David McNaughton and Piers Rawling claim that these criticisms are unfounded. They trace the intuitionist's beginnings from the work of W.D. Ross, and show that Ross's work is much more systematic and sophisticated than most philosophers suppose.

In a full-blown intuitionism, like that defended by McNaughton and Rawling, intuition takes on a narrower sense than it has for McMahan. For unlike McMahan, McNaughton and Rawling claim that intuitions give us a priori knowledge of self-evident moral principles that are distinct in kind from other claims about reality. To say they are self-evident does not mean they are immediately obvious, or understandable by the most uneducated and morally unenlightened dolt. Rather, self-evident truths can be discerned only by intelligent and experienced people, who have appropriately reflected on moral matters. The correct moral theory is not a monistic theory which tries to derive our duties from a single exceptionless general principle, but a pluralistic theory. Ross lists a number of "*prima facie* duties" – in more modern terminology, moral reasons – which can conflict and have to be weighed against each other to establish what we ought to do. There is no algorithm or mechanical procedure for determining what wins out in this weighing. McNaughton and Rawling end by exploring the intricate relationship between Rossian intuitionism and the more recent theory known as *particularism*.

In the following essay, Thomas E. Hill Jr tries to disentangle the central elements of "Kantianism" from the more peripheral ones. Kant's moral writings are standardly studied as major works in the history of moral philosophy, but there was a period in which his moral theory was not taken to be a live option. More recently, however, Kantianism has reemerged as a prominent moral theory, in no small measure because of a string of commentators who have infused new life into his thought. The problem, Hill notes, is that Kant has often been interpreted as advocating rather radical ideas, including the ideas that (1) empirical evidence is irrelevant to moral deliberation and that (2) the only actions of moral worth are those done from duty and against the agent's inclinations. To many, such ideas seem psychologically untenable, epistemologically uninformed, and morally odious.

Perhaps Kant holds these views – although Hill is far from sure that he does. Nonetheless, they are not the core of Kant's thought. And that core should not be lost because of squabbles over Kant's more radical views. These core ideas survive intact. They are significant developments in moral thought: the important but limited role of the a priori method, the basic contours of Kant's account of duty, the nature of the Categorical Imperative, and the idea that his account of duty presupposes that we are autonomous agents.

Kant is also a pivotal figure in the history of "Contractarianism" (also called contractualism). Geoffrey Sayre-McCord chronicles the development of contractarianism from its ancient beginnings to the current day when it has again become popular through philosophers like John Rawls. In its beginnings the contractarian

approach was an attempt to justify the state and political legitimacy, and the appeal was to an actual contract between those governed by the state. As it soon became apparent that it was difficult to identify an actual contract, the appeal shifted to a *hypothetical* contract under more or less idealized conditions. But if the contractual circumstances become hypothetical, the distinctive contractarian idea that the justification for rules and institutions flows from a consent to them tends to be replaced by the idea that justification flows from the reasons there are to give consent. These can be utilitarian, for instance, and then the resulting theory basically becomes a sort of utilitarianism.

Sayre-McCord suggests that there are two main traditions of contractarianism: a Hobbesian version, according to which the contracting parties can be selfish; and a Kantian version, according to which they are constrained by some measure of morality. The chief problem of the Kantian version is that this reliance upon a measure of morality independent of contractarianism makes its contractarianism less than thoroughgoing. By contrast, the Hobbesian version offers to account for *all* of morality, but it is a morality that differs radically from commonsense morality. Sayre-McCord favors a Humean version which seeks to explain why and how morality would have naturally emerged in human society; how the sorts of creatures we are would develop the kinds of practices and employ the evaluative concepts that we do. Sayre-McCord surmises that such explanations of our practices can also generate justifications of them.

In the next essay, L.W. Sumner explores the leading role that "Rights" play in many deontological theories. Rights set constraints on attempts to maximize the social good, and they thereby safeguard individuals against the intrusive interests of society or other individuals. Rights focus on their possessors – on the agents whose interests they protect – rather than the agents who must respect those rights. This focus, Sumner argues, gives theories that accommodate rights a significance absent from theories without them.

Sumner provides a scheme for classifying rights, and tries to show precisely what rights require and what they protect. He condemns the popular tendency to assert rights to everything we want; a tendency that has led to a senseless proliferation of, and thereby a diminution of the significance of, rights. Rights are not moral toys we construct at will; they require a theory that explains and grounds them. The best ground for rights is provided – somewhat surprisingly – by a goal-based consequentialist theory rather than a deontological one.

Jan Narveson agrees with Sumner that rights are an important moral currency. But, unlike Sumner, "Libertarianism" holds that there is only one moral right – the right to liberty. And that right is (virtually) inviolable. Thus, libertarians share the basic presupposition of other nonconsequentialists, namely, that we should not override individuals' rights to maximize the good. But libertarians think that most nonconsequentialists have too broad an understanding of right and wrong and, therefore, are too willing to override constraints against violating rights.

The central notion of libertarianism is self-ownership. The proper moral order has one aim: to protect individuals' rights to themselves. Coercion is justified

only to control actions aggressing against others. Nonetheless, Narveson insists, libertarianism need not be seen as a selfish, narrowly individualistic theory. Libertarians can establish and support communities that urge their members to help others in need. Indeed, libertarians will not be averse to saying that each of us has a duty to provide mutual aid, as long as we understand that this is not an enforceable duty.

One of the most notable theoretical developments of the past three decades has been the emergence – or reemergence – of an alternative that challenges, and dramatically diverges from, consequentialism and deontology: "Virtue Ethics." Its roots go back to Plato and Aristotle. Still, after several centuries of oblivion it has only recently reappeared on the theoretical stage in the Western world, though in Asia it has been continuously alive since antiquity, in particular in the shape of Confucianism. Although some people might suspect that the gap between virtue ethics and the standard alternatives of consequentialism and deontology is slight, Michael Slote claims the theories are different to the core. Whereas both consequentialism and deontology treat deontic concepts of "ought," "right," "duty," and "obligation" as the central moral concepts, virtue theorists hold aretaic notions like "excellence" and "admirable" as key. More specifically, virtue theorists are especially concerned about inner states of character and motivation. Although deontologists or consequentialists may also be concerned about character, their concern is derivative: character matters only because it makes people more likely to promote the good or to follow moral rules. Consider, for instance, Frey's argument that character plays a central role in the proper understanding of act-utilitarianism. In contrast, virtue theorists see virtue as primary and deontic notions as derivative.

Slote distinguishes between a *rationalist* form of virtue ethics, which he sees Aristotle as advocating, and a *sentimentalist* form, which he finds in Hume. In Slote's view, Aristotelian rationalism, with its stress on the superiority of the morally wise person, has troubles fitting in with the current political ideals of democracy and toleration. Sentimentalist virtue ethics has greater resources of accommodating such ideals by means of the virtues of empathy and humility. On the other hand, sentimentalism is harder put than rationalism to provide a ground for the objectivity that we are inclined to attribute to morality. So, Slote concludes that contemporary virtue theorists, whichever their persuasion, have a pretty full agenda.

"Capability Ethics" is a recent addition to the ethical landscape. Ingrid Robeyns argues in her essay that it has not yet been developed into a full moral theory. It is best seen as offering an alternative to the concept of well-being or welfare which occupies a central place in many ethical theories, in particular utilitarianism. Human "capabilities" refer to the combination of internal and external conditions which are necessary for humans to be or to do something, in particular something that they regard as valuable. The two main advocates of capability ethics are Martha Nussbaum and Amartya Sen, and Robeyns spends a good deal of her essay detailing differences between their conceptions of capabilities. Further questions that

Robeyns discusses are to what extent the capability approach offers a theory of social justice, what the relations are between capabilities and human rights, and whether the capability approach should be regarded as deontological or consequentialist theory. She concludes by emphasizing respects in which capability ethics needs to be further developed to amount to a full theory of morality.

In the penultimate essay – "Feminist Ethics" – Alison M. Jaggar argues that all Western ethical theories have consistently devalued women. This devaluation has been captured and rationalized in these theories' central concepts and reasoning. Even after ethical theorists acknowledged the basic equality of men and women, they still refused to criticize or challenge the myriad ways in which women have been and continue to be disadvantaged, or the ways in which their theories support that disadvantage.

How are these disparities to be remedied? Minimally, standard ethical categories must be expanded to give due attention to significant issues affecting women. Some women have also proposed that women's ethical experience should be explicitly given a central role in ethical theory. Most notably this is seen in the development of a care ethic. Although Jaggar thinks the care ethic has been a significant development, in part because it exposes some central flaws in modern ethical theory, the theory is inadequate. The care perspective must be supplemented by a capabilities approach that first arose in, and now informs, debates about third world development.

The view advanced by William R. Schroeder in the last essay, "Continental Ethics," notably differs from that taken by most essayists in this book: essayists who are representatives of the analytic (Anglo-American) tradition. According to Schroeder, Continental thinkers are suspicious of conventional morality, much more so than analytic moral philosophers generally are. In spite of this difference, many current developments in analytic ethical theory – especially the search for alternatives to consequentialism and deontology and the renewed interest in moral realism – have their roots in Continental thought.

Perhaps the most notable difference between Continental and analytic ethics, Schroeder claims, is the Continentalist's emphasis on personal growth, authenticity, and creativity. He shows how these ideas were developed in the work of several pivotal Continental thinkers: Hegel, Nietzsche, Scheler, Sartre, and Levinas. Schroeder then explores ongoing questions about whether values are found or created by humans and questions about the priority of liberty. He challenges what he claims is a guiding assumption of most analytic ethical theories: that the main job of ethics is to suppress people's basic selfishness.

Prospects for Future Ethical Theory

This book does not explicitly discuss the history of ethics, although some elements of that history are evident in the discussions of individual authors. Nor does this

book pretend to discuss all the relevant issues or to provide a final solution to the questions that have plagued philosophers for thousands of years. Its aim is more modest: to provide a way station on the long and distinguished journey of ethical theory. Our hope is that it not only reliably captures the current state of debate but also will prompt further productive work in ethics.

References

Ayer, A.J. (1946/1952) *Language, Truth, and Logic*, 2nd edn, New York: Dover Publications.

Charles, S. (1944) *Ethics and Language*, Yale University Press.

Darwall, S. (1998) *Philosophical Ethics*, Boulder, CO: Westview Press.

Hare, R.M. (1952) *The Language of Morals*, Oxford University Press.

Kagan, S. (1998) *Normative Ethics*, Boulder, CO: Westview Press.

Part I

Metaethics and Moral Epistemology

Chapter 1

Moral Realism

Michael Smith

In the past thirty years or so, the debate over moral realism has become a major focus of philosophical activity. Unfortunately, however, as a glance at the enormous body of literature generated by the debate makes clear, there is still no consensus as to what, precisely, it would take to be a moral realist (Sayre-McCord 1988a). My aims in this essay are thus twofold: first, to clarify what is at stake in the debate over realism, and, second, to explain why, as it seems to me, the realist's stance is more plausible than the alternatives.

Moral Realism vs Nihilism vs Expressivism

What do moral realists believe? The standard answer is that they believe two things. First, they believe that the sentences we use when we make moral claims – sentences like "Torturing babies is wrong" and "Keeping promises is obligatory" – are capable of being either true or false, and, second, they believe that some such sentences are true. Moral realism thus contrasts with two quite distinct kinds of view.

The first view shares realism's first commitment, but rejects the second. According to this first alternative, when we make claims about acts being obligatory, right, and wrong we intend thereby to make claims about the way the world is – we intend to say something capable of being either true or false – but none of these sentences are true. When we engage in moral talk we presuppose that obligatoriness, rightness, and wrongness are features that acts could possess, but we are in error. There are no such features for acts to possess. This view generally goes under the name of nihilism or the error theory (Mackie 1977; Joyce 2001).

The Blackwell Guide to Ethical Theory, Second Edition. Edited by Hugh LaFollette and Ingmar Persson.

The second more radical view shares neither commitment. According to this view, the sentences we use when we make moral claims are not used with the intention of saying something that is capable of being either true or false. We do not use them in an attempt to make claims about the way the world is. By contrast with nihilism, we therefore do not presuppose that obligatoriness, rightness, and wrongness are features that acts could possess. Rather we use moral sentences to express our feelings about acts, people, states of the world, and the like. When we say "Torturing babies is wrong" it is as if we were saying "Boo for torturing babies!" This view generally goes under the name of noncognitivism or expressivism or projectivism (Hare 1952; Gibbard 1990; Blackburn 1994).

Expressivism and nihilism share a conception of the world as value-free and so devoid of any moral nature. However, they differ in a crucial respect as well. Because nihilism insists that moral thought and talk presuppose that obligatoriness, rightness, and wrongness are features of acts, it sees the value-free nature of the world as something that demands a reform of moral practice. We cannot continue to assert falsehoods once we know them to be false, but must rather refrain from asserting them at all, or else justify the pretense that the falsehoods are true. Moral thought and talk thus have the same status as religious thought and talk once we become convinced atheists, at which point we must either stop going to church altogether, or else continue to go but do so for nonreligious reasons such as the love of the community or the music. By contrast, expressivism holds that the value-free nature of the world has no such consequence. It holds that moral thought and talk can proceed perfectly happily in the knowledge that the world is value-free because, in making moral claims, we never presupposed otherwise.

The upshot is that there are therefore two fundamental – if rather abstract and general – questions that need to be answered to resolve the moral realism debate. The first is whether sentences that ascribe obligatoriness, rightness, and wrongness to actions are capable of being true or false – if we answer "yes" to this question then we thereby refute expressivism – and the second, which presupposes an affirmative answer to the first, is whether any sentences ascribing obligatoriness, rightness, and wrongness to actions are true. If we answer "yes" to this second question then we thereby eliminate the nihilist option as well. Answering "yes" to these two questions commits us to the truth of moral realism.

An Initial Difficulty

So described, moral realism looks to be a very demanding doctrine. It can go wrong in two distinct ways. Perhaps it wrongly supposes that sentences ascribing obligatoriness, rightness, and wrongness to actions are capable of truth and falsehood, or, granting that it is right about that, perhaps it wrongly supposes that some of these sentences really are true. However, as we will see, the real danger

is that moral realism, so understood, is insufficiently demanding. As characterized, it may be too easy to be a moral realist.

The distinctive feature of the two abstract and general questions just asked is that they each involve semantic ascent; that is, they each speak of a feature that must be possessed by the sentences we use when we make moral claims, or a relation that must obtain between these sentences and the world. But the fact that they each involve semantic ascent poses an initial difficulty. If a commitment to the truth of moral realism comes by answering "yes" to these two abstract and general questions, then it looks as if such commitment might come cheaply, at least to competent speakers of English who have any moral commitments at all. Let me illustrate the difficulty.

Like most people reading this essay, I have various moral commitments. For example, I am quite confident that torturing babies is wrong. As a competent speaker of English, I am therefore willing to say so by using the English sentence "Torturing babies is wrong." Imagine me saying this out loud:

Torturing babies is wrong.

Moreover, as a competent speaker of English, I am also willing to say so not just by *using* this sentence of English but also by *mentioning* it. Imagine me saying this out loud:

"Torturing babies is wrong" is true.

Or even

"Torturing babies is wrong" is really true.

This is because, in common parlance, mentioning this sentence and saying of it that it is true is simply an alternative way of saying what I could have said by using the sentence. "'Torturing babies is wrong' is true" and "'Torturing babies is wrong' is really true" are simply long-winded ways of saying that torturing babies is wrong – ways that involve semantic ascent.

Given the initial characterization of what it takes to be a moral realist, it therefore seems to follow that I am a moral realist. After all, since I willingly assert the truth of "Torturing babies is wrong" it follows that I think that the sentences I use when I make moral claims – sentences like "Torturing babies is wrong" – are both capable of being true or false and that some of these sentences really are true . . . oh dear. Something has clearly gone wrong. Perhaps a commitment to moral realism follows from the mere fact that I have moral commitments, together with the fact that I am a competent speaker of English, but it seems very unlikely. But what exactly has gone wrong?

An obvious suggestion is that the surface grammar of moral sentences is potentially misleading, masking some deeper metaphysical fact. Though we *say* that these

sentences are true and false, this is loose talk. What moral realists really believe, the suggestion might be, is that the sentences we use when we make moral claims are capable of being true or false *strictly speaking*. Expressivists, by contrast, hold that moral claims are only capable of being true or false *loosely speaking*. Everything thus turns on what it is to speak strictly, as opposed to loosely, when we say of sentences that they are true or false.

Minimalism

What do the words "true" and "false" mean strictly speaking? One very popular view nowadays is *minimalism about truth* (Horwich 1990; Wright 1992). According to this view, the role of the words "true" and "false" in our language is simply to enable us to register our agreement and disagreement with what people say without going to the trouble of using all the words that they used to say it.

For example, suppose A says "Snow is white, and grass is green, and roses are red, and violets are blue," and that B wants to register agreement. If the word "true" was not a part of our language then, in order to do so, B would have to quote what A said and then disquote. B would have to say "A said 'snow is white' and snow is white, and A said 'grass is green' and grass is green, and A said 'roses are red' and roses are red, and A said 'violets are blue' and violets are blue." But that requires B to use more than twice the number of words that A used. The role of the word "true," according to the minimalist, is simply to allow B to register agreement more efficiently. Because we have the word "true" in our language, B can quantify over all of the things that A said and then say, all at once, "Everything B said is true."

The upshot, according to minimalism, is that all there is to say about the meaning of the words "true" and "false," strictly speaking, is precisely what we said when noting the initial difficulty. All there is to know about the meaning of the word "true" is that, when "s" is a meaningful sentence of English, and when "'s' is true" is also a meaningful sentence of English, someone who says "'s' is true" could just as well have disquoted and said instead "s." When you mention or quote an English sentence and meaningfully append "is true" to it, this is just another way of saying what could have been said by using or disquoting that English sentence. Minimalism about truth thus suggests that when I say "'Torturing babies is wrong' is true," rather than "Torturing babies is wrong," I *am* speaking strictly, for I thereby register the appropriateness of disquotation.

Accordingly, it seems to me that we should therefore put a very first realist option on the table. Minimal moral realists believe three things. First, they believe that the sentences we use when we say that actions are right and wrong are true or false strictly speaking, rather than merely loosely speaking; second, they believe that some of these sentences really are true; and third, they believe that, strictly speaking, the meanings of the words "true" and "false" are fully explained by the

minimalist's story. Minimal moral realism is a very cheap doctrine indeed: if you accept the minimalist's story about truth then, if you have any moral commitments at all, you are a moral realist – or, at any rate, you are a minimal moral realist. Nihilism and expressivism are eliminated in one fell swoop. The obvious questions to ask are whether we should all be minimal moral realists and, if so, whether nihilism and expressivism really are so easily eliminated.

Why Minimalism Does Not Really Make a Difference

Minimalists about truth tell us that all there is to know about the meaning of the word "true" is that, when "s" is a meaningful sentence of English, and when "'s' is true" is also a meaningful sentence of English, someone who says "'s' is true" could just as well have disquoted and said instead "s." But this story – at least in the form in which it has just been told – buries an extra, crucially important, piece of information about truth, for it fails to tell us the conditions that need to be satisfied by "s" in order for "'s' is true" to be a meaningful sentence of English. In other words, it fails to tell us what it is about a sentence that is capable of truth and falsehood that *makes* it capable of truth and falsehood. Let me spell out this problem in greater detail (Jackson, Oppy, and Smith 1994).

Everyone agrees that "Snow is white" and "'Snow is white' is true" are both meaningful sentences of English. Moreover, everyone also agrees that though "Hooray for the Chicago Bulls!" is a meaningful sentence of English, "'Hooray for the Chicago Bulls!' is true" is not. But why is there this difference between the two sentences? What do the meaningful strings of English words that are truth-apt have in common that they do not have in common with those strings of English words that are non-truth-apt? What feature of the truth-apt sentences of English makes them truth-apt? Minimalism about truth, as so far characterized, does not provide an answer. Yet surely an answer to this question is part of what we need to know, when we know all there is to know about the meaning of the word "true."

Minimalists about truth typically insist that they can provide a suitably minimal answer to this question (Wright 1992; Horwich 1993). Consider three strings of English words: "Snow is white," "Torturing babies is wrong," and "Hooray for the Chicago Bulls!" The standard minimalist suggestion is that the first two strings of English words are truth-apt, and the third is not, because of a purely *syntactic* feature that they possess and the third lacks. The first two strings of English words, they suggest, are of an appropriate grammatical type to figure in a whole array of contexts: the antecedents of conditionals (for example, "If snow is white, then it is the same color as writing paper" and "If torturing babies is wrong then I will support the existence of a law against it" are both well-formed sentences), propositional attitude contexts ("John believes that snow is white" and "John believes that torturing babies is wrong" are both well-formed sentences), and so on and

so forth. But the third sentence, by contrast, is not of the appropriate grammatical type to figure in these contexts (neither "If hooray for the Chicago Bulls then I will get tickets to see them play next season" nor "John believes that hooray for the Chicago Bulls" are well-formed sentences). It is this syntactic feature of the first two sentences that, according to the minimalists, makes it appropriate for them to figure in "'——' is true" contexts, and it is the fact that the third lacks this feature that makes it incapable of figuring in such contexts – so, at any rate, minimalists typically argue.

However, for reasons Lewis Carroll made plain in his wonderful poem "Jabberwocky," this minimalist account of truth-aptitude is unsatisfactory (Carroll 1872/1998). "'Twas brillig, and the slithy toves did gyre and gimble in the wabe" looks like a conjunction of sentences which, syntactically, are of the appropriate grammatical type to figure in the antecedents of conditionals (thus, for example, "If the toves are gyring and gimbling in the wabe then I will watch them" looks for all the world to be a well-formed sentence), to be embedded in propositional attitude contexts ("I believe that the toves are gyring and gimbling" looks to be a well-formed sentence), and so on. Indeed, it looks like these sentences can have "true" predicated of them ("'The toves are gyring and gimbling in the wabe' is true" looks to be a well-formed sentence). But it does not follow that these sentences are truth-apt. Indeed, we know that they are not truth-apt because, notwithstanding their syntax, they are nonsense sentences – sentences without any meaning whatsoever. They are therefore incapable of being either true or false. The idea that mere syntax is sufficient to establish truth-aptitude is thus absurd.

We must therefore ask what a sentence with the right syntax must have added to it in order to make it truth-apt. For example, what feature would Carroll's sentence "the slithy toves did gyre and gimble in the wabe" have to have added to it, in order to make it truth-apt? The obvious answer to this question is that the sentence would have to be meaningful, rather than nonsense, and for this to be the case the constituent words in the sentence – words like "tove," "gyre," "gimble," and "wabe" – would have to be associated with patterns of usage that make it plain what information about the world people who use the words in those ways intend to convey when they use them. And this, it might be said, is exactly what we had in mind earlier when we said that certain sentences are capable of being true or false strictly speaking. They are capable of being true of false strictly speaking when their meanings are explained by reference to states of the world.

If this is right then it might be thought to follow that truth-aptitude is not an entirely minimal matter. If sentences which are truth-apt have to be sentences that could, in principle at least, be used to convey information, then, the thought might be, they must be sentences that could, in principle, be used to give the content of people's beliefs where the role of these beliefs is in turn explained by the relationship in which they stand to the states of the world that the information is about. (In what follows, I will ignore the complication entailed in avoiding Moorean problems with meaningful sentences like "I have no beliefs" by

determining that they must be sentences that are suitably related by some grammatical transformation to sentences that could, in principle, be used to give the contents of people's beliefs.) But since it is a substantive fact about a sentence that its constituent words are associated with patterns of usage that allow them to convey information about particular aspects of the world, and since we discover this substantive fact when we discover which beliefs the sentence can be used to express, it is therefore this substantive fact about a sentence that we need to discover in order to establish that it is truth-apt.

The minimalist might reply that this all assumes that the *only* way to explain the meanings of sentences with the syntactic features of a truth-apt sentence is by reference to the states of the world about which those sentences can be used to convey information. But, the minimalist might continue, this assumption is false. We might instead associate uses of those sentences with the expression of certain nonbelief states (Dreier 2004). Importantly, the minimalist might add, this would not be to deny that those meaningful sentences are truth-apt; nor would it be to deny they can be used to express beliefs, as both of these facts about such sentences would be guaranteed by the fact that they have the syntactic features that they do. It would simply be to acknowledge that there are two very different ways in which we might explain the meanings of truth-apt sentences. In the case of sentences about moral matters, which are truth-apt, we explain their meaning in the alternative way, that is, by reference to the desires that constitute people's beliefs about the moral matters that those sentences can be used to express.

As I hope is clear, however, this form of minimal moral realism turns out to be just expressivism by another name. The only difference is that, whereas expressivists deny that moral sentences are truth-apt strictly speaking and yet admit they may nonetheless be truth-apt loosely speaking, this form of minimal moral realism holds that there are two ways to explain meaning of a truth-apt sentence. One way is to tell a story about the states of the world that that sentence can be used to convey information about, and which is thus to tell the story of the meanings of moral sentences according to which those sentences are what the expressivist calls truth-apt *strictly speaking*. The other is to tell an expressivist story. In other words, it is to tell the story of the meanings of moral sentences according to which those sentences are what the expressivist calls truth-apt merely *loosely speaking*.

The upshot is that the detour via minimal moral realism does not really make any difference. For either the minimal moral realist is committed to explaining the meanings of moral terms in terms of the features of the world about which they can be used to convey information – more on this presently – or the minimal moral realist is committed to explaining the meanings of moral terms in terms of the desires that the use of such terms expresses. The latter version of minimal moral realism is thus also refuted if we can show that the meanings of moral sentences can be given in terms of the states of the world about which they can be used to convey information, for doing that would establish that such sentences are

truth-apt strictly speaking. The detour via minimalism thus leaves everything exactly where it was beforehand.

Expressivism and Internalism

We now know what would have to be the case for sentences like "Torturing babies is wrong" and "Keeping promises is obligatory" to be capable of being true or false strictly speaking. The words contained in these sentences – words like "right" and "wrong" – would have to be associated with patterns of usage that make it plain what information about the world people's use of them is intended to convey. The question to ask is, therefore, whether the patterns of usage associated with the words "right" and "wrong" have this striking feature. Can we give an account of the information about the world that the use of such words is intended to convey? Many people argue that we cannot.

They begin by noting the very striking fact that people's moral views tell us something about their dispositions to action. For example, it would be extremely puzzling if, having announced your firm conviction that it would be wrong to fail to give money to Oxfam, you then claimed utter indifference to actually giving money to Oxfam when the opportunity arose. Perhaps your indifference could be explained away. Depression and weakness of will can, after all, sap our desire to do what we think is right. But, absent some such explanation, it seems that your indifference would give the lie to your announced conviction. It would reveal you to be a hypocrite. This is why, when it comes to expressing moral views, actions speak louder than words.

This striking fact is called the *internalism constraint* (Hare 1952: ch. 1; Blackburn 1984: 187–9). According to internalists, there is an internal or necessary connection between the moral judgments we make and our motivations. If true, internalism places a constraint on the proper use of moral sentences. It tells us that it is a constraint on the proper use of "Torturing babies is wrong" that someone who sincerely utters it is averse to torturing babies, at least other things being equal (in other words, absent depression, weakness of will, and the like). Likewise, it tells us that it is a constraint on the proper use of "Keeping promises is obligatory" that someone who sincerely utters it desires to keep promises, at least other things being equal.

Expressivists seize on the truth of internalism and ask the obvious question, how could the proper use of moral sentences be constrained by the truth of internalism if they could be used to give the contents of people's beliefs where these beliefs are not in turn constituted by people's desires? After all, when we consider sentences that can uncontroversially be used to give the contents of such beliefs – sentences like "Snow is white," "London is north of Paris," "If you waste your time in school then you will diminish your life prospects in the future," and the like – we note that they can all be used by people perfectly sincerely *no matter*

what pattern of desire and aversion these people have. Why should there be any difference with moral sentences?

For example, it is not a constraint on the proper use of the sentence "If you waste your time in school then you will diminish your life prospects in the future" that someone who sincerely utters this sentence desires people to waste their time in school, or is averse to people wasting their time in school, or is indifferent to people wasting their time in school. The belief that wasting your time in school will diminish your life prospects in the future can quite happily coexist with any of these attitudes. Moreover, we find this same pattern of possibilities when we consider any other sentence that can uncontroversially be used to express the content of people's beliefs where those beliefs are not constituted by desires. So, expressivists ask, why do we not find that same pattern of possibilities in the case of the sentences "Torturing babies is wrong" and "Keeping promises is obligatory" if they too express beliefs? Why cannot the belief that keeping promises is obligatory coexist perfectly happily with the desire to keep promises, aversion to keeping promises, and indifference to keeping promises? What is it about this belief, if there is any such belief, that makes it require the presence of a desire to keep promises?

The answer, according to expressivists, is that we do not find that same pattern of possibilities because the sentences "Torturing babies is wrong" and "Keeping promises is obligatory" can only be used to give the content of a belief that is itself, in turn, constituted by a desire. It is not by reference to the states of the world about which they can be used to give information that we explain the meanings of moral sentences. Instead we explain their meanings by reference to the desires that constitute the beliefs the contents of which such sentences can be used to express. The proper role of "Torturing babies is wrong" is to express aversion to torturing babies, and the proper use of "Keeping promises is obligatory" is to express the desire to keep promises.

Accordingly, expressivists hold that moral sentences, when properly understood, are very similar to other uncontroversially non-truth-apt sentences – sentences like "Hooray for the Chicago Bulls!" The latter sentence is non-truth-apt, strictly speaking, not because of its surface syntax, according to expressivism, but, instead, because it is properly used to express a pro-attitude toward the Chicago Bulls, a want or a desire of some kind, rather than any belief about the Chicago Bulls. Likewise, sentences about actions being right, wrong, and obligatory are non-truth-apt, strictly speaking, because they are properly used to express desires and aversions with regard to those actions. When we say of moral sentences that they are true or false, we are therefore at best speaking loosely, not strictly. Strictly speaking, moral sentences cannot be true or false. They cannot be used to convey information about the way anything is. There are no moral beliefs for anyone to express where the truth of those beliefs would be a matter of truth and falsehood strictly speaking.

Expressivists are thus best seen as offering a challenge to both moral realists and nihilists. They challenge both these theorists to explain how the use of moral

sentences could be constrained by the truth of internalism if their proper use was to convey information. What information is it such that, in order to possess that information, you have to have certain desires or aversions? Expressivists intend this question to be rhetorical. Even so, many opponents of expressivism – both moral realists and nihilists – have tried to answer the challenge. But in order properly to answer the challenge we can now see that they must do more than simply stamp their feet and insist that moral sentences *can* be used to give the contents of beliefs. They must specify, in precise terms, what states of the world these beliefs are about. Let us focus in on a particular example of an attempt to do just that.

Naturalistic Moral Realism

As I said at the outset, I am quite confident that torturing babies is wrong, and I am quite willing to say so by using the English sentence "Torturing babies is wrong." But if my use of this sentence expresses some belief I have about torturing babies, where "Torturing babies is wrong" is true strictly speaking, then it is fair and reasonable to ask what the content of that belief is. What feature of the world would make it true that torturing babies is wrong? It might be thought that we could just give the glib answer: torturing babies would have to have the feature of being wrong. But it turns out that I have to be able to say much more than this.

If there is some feature of torturing babies that makes it true that torturing babies is wrong, then, in giving an account of that feature, we are constrained by our conception of the world in which we live. This means, in turn, that we are constrained by the truth of *naturalism*, the view that the world is amenable to study through empirical science. This is because, given the success of the empirical sciences in providing explanations of various aspects of the world, it is extremely plausible to hold that the world is *entirely* amenable to study through the empirical sciences. Naturalism accordingly entails that the only features we have any reason to believe objects have are one and all naturalistic features – features which are themselves posits, or composites of posits, of empirical science. The upshot is therefore that, if any form of moral realism is true at all, then it must be a form of *naturalistic moral realism* (Railton 1993a, 1993b).

Naturalistic moral realism holds not only that some of the sentences that we use to make moral claims are capable of being true and false, strictly speaking (from hereon I will omit this qualification, as I will only speak of truth and falsehood strictly speaking), and that some are true, but also that what makes the true ones true are naturalistic features of the world: features amenable to understanding in scientific terms. If moral features exist at all then, given the truth of naturalism, it follows that they too must be features that can be discovered either directly by observation, or by inference from observational information. Moral beliefs must

therefore have naturalistic contents, for only so could they be made true by naturalistic features of the world.

We can now ask a more specific version of the question we asked earlier. If, as naturalistic moral realists suppose, the sentence "Torturing babies is wrong" can be used to give the content of a belief, then what *naturalistic feature* does someone with this belief thereby believe torturing babies to have? This is the question naturalistic moral realists must answer. Moreover, they must answer this question by appealing to some constraint on the way in which we use moral words. There must be some constraint on our use of moral words that makes these words apt to pick out a natural feature of acts.

It is not difficult to see what that constraint might be. By all accounts it is a conceptual truth that the moral features of acts *supervene* on their naturalistic features: two acts which are identical in all of their natural features must be alike in their moral features as well. It thus follows that if we acknowledge that a particular act is right but insist that another act, exactly the same in every naturalistic respect, is not right, then we thereby misuse the word "right," likewise for "obligatory" and "wrong." When we apply "obligatory," "right," and "wrong" to acts, we are thus constrained to do so in the belief that the acts in question have some naturalistic feature that *warrants* the ascription of "obligatory," "right," or "wrong." This is the *supervenience constraint* (Jackson 1997).

The fact that we are constrained to use moral words in the way just described – that is, to ascribe moral features in virtue of naturalistic features – requires an explanation, however. Why cannot we say that acts are alike in all naturalistic respects and yet differ morally? Why cannot moral features float free of naturalistic features altogether? The answer favored by naturalistic moral realists is that this is because moral features *are* natural features.

If this is agreed, then the only question left to answer is *which* natural features warrant the ascription of various moral features to acts. Once we know the answer to this question then, according to naturalistic moral realists, we should simply conclude that moral features are those natural features. For example, if the naturalistic feature of acts that warrants the ascription of rightness turns out to be utility maximization, then, according to naturalistic moral realists, rightness *is* utility maximization. The answer to the question "Which naturalistic feature does someone with the belief that torturing babies is wrong thereby believe torturing babies to have?" will then turn out to be the feature of failing to maximize utility.

The Open Question Argument

Elegant though this suggestion might be, it faces a serious objection. The objection was first put forward by G.E. Moore (1903). To stick with our example, Moore agreed that acts of utility maximization might always have the feature of being right, but he insisted that we resist concluding that the properties of

maximizing utility and being right are the same property. They are, he insisted, quite distinct properties. The argument he gave for this conclusion is his famous Open Question Argument.

Suppose, for reductio, that rightness and utility maximization were indeed one and the same feature of acts. Then, according to Moore, it would follow that "rightness" and "utility maximization" are analytically equivalent. But it is quite clear that "rightness" and "utility maximization" are not analytically equivalent. After all, if they were analytically equivalent then the question "This act maximizes utility, but is it right?" would have to be one whose answer is immediately obvious to anyone who understands the meanings of the words used, being equivalent to the question "This act maximizes utility, but does it maximize utility?" However, as a moment's reflection reveals, the questions clearly are not equivalent. We can, without self-contradiction, agree that an act maximizes utility but deny that it is right. It therefore follows that the question "This act maximizes utility, but is it right?" is not a *closed question* – one whose answer is immediately obvious to anyone who understands the meanings of the words – but is rather an *open question* – one whose answer is open to reasoned argument. "Rightness" and "utility maximization" are thus not analytic equivalents. They do not pick out the same feature.

If the Open Question Argument is sound then it delivers a very strong conclusion indeed. For, as Moore pointed out, it does not seem to matter which of the various natural features of acts we consider. It would always be open to reasoned argument whether an act with any of the various natural features we might care to consider is right. It is, for example, an open question whether an act of keeping a promise is right, an open question whether an act which advances my own well-being is right, an open question whether acts that I desire to perform are right, and so we could go on and on. No matter which natural features we choose, it seems that it is never obvious whether an act with such features is right. It is always open to reasoned argument. So, if sound, the Open Question Argument seems to show that rightness is not identical with any natural feature of acts at all. It thus constitutes a decisive refutation of naturalistic moral realism. But if we should not accept naturalistic moral realism, then which theory should we accept instead?

Nonnaturalistic Moral Realism

Moore himself thought, on the basis of the Open Question Argument, that we should reject naturalism altogether and admit a realm of extra, *sui generis*, non-natural properties into our ontology. The property in question was *goodness*, and Moore thought that we could then appeal to goodness to explain obligatoriness, rightness, and wrongness.

Moore thus embraced *nonnaturalistic moral realism*. He believed not only that some of the sentences we use to make moral claims are capable of being true and

false, and that some are true, but also that what makes the true ones true are nonnaturalistic states of the world – states that elude understanding in scientific terms, namely, the distribution of goodness. Beliefs about which acts are right and wrong are thus beliefs about the nonnatural features possessed by the states of affairs that are caused by acts. Moreover, according to Moore, some such beliefs represent the world to be the way it really is. Moore was no naturalist.

However, the problems with Moore's own version of nonnaturalistic moral realism are evident and overwhelming (Blackburn 1984: ch. 6). The first problem is that it must explain how we come by knowledge of goodness. Unsurprisingly, however, Moore had no explanation. He could hardly claim that we come by knowledge of goodness via observation, for any property knowable in that way is, by definition, naturalistic. But nor could he claim that we come by knowledge of goodness via inference from any of the naturalistic features of acts, for that is precisely what the Open Question Argument (allegedly) shows to be impossible. The only options left seem to multiply the mysteries. For example, we might suppose that there is some nonempirical sort of observation: a sort of spooky sixth sense that allows us to detect the presence of goodness. But as soon as the idea is stated it is plain that it is, in reality, too absurd to take seriously.

The second problem for nonnaturalistic moral realism is that it must explain why there are not possible worlds in which goodness floats free of the natural properties with which it is coinstantiated in actuality. Perhaps in actuality goodness is coinstantiated with the presence of (say) happiness. But if goodness is a distinct feature, then why are there not possible worlds in which goodness is not coinstantiated with happiness? Again, Moore had no explanation of why this possibility is ruled out. If nonnatural properties are distinct from natural properties then, it seems, we should be able to pull them apart modally. Given that moral properties supervene on natural properties, however, it follows that we cannot pull them apart modally. Moore was thus forced to view the supervenience of the nonnatural on the natural as a brute mystery.

The third problem is that it is unclear why anyone should care about the existence of a nonnatural feature. We ordinarily suppose that morality is a matter of great practical significance. Expressivism is in many ways tailor-made to capture this idea. According to the expressivists, morality is a matter of great practical significance because in making moral judgments we thereby express our deepest cares and concerns. Nonnaturalists, by contrast, hold that when we make moral judgments we express our beliefs about the nonnatural properties possessed by things. But if that is all we are doing, why do we care so deeply about morality? What is the attraction of the nonnatural property of goodness that makes so many of us prefer states of affairs with this nonnatural property to those that lack the property? Moore provides us with no answer.

Modern versions of nonnaturalistic moral realism fare somewhat better than Moore's own version (Scanlon 1998; Parfit 2011). They suppose that goodness is not itself a nonnatural feature, but rather that it is the property of having some natural feature that provides everyone with a reason to desire that that natural

feature is instantiated. Consider again possible worlds in which a great deal of happiness is instantiated. According to the modern form of nonnaturalistic moral realism, these possible worlds are good to the extent that they have a natural feature, happiness, which provides everyone with a reason to desire that that feature is instantiated. This is still a version of nonnaturalistic moral realism because the reason relation itself is conceived of in entirely nonnaturalistic terms. In other words, according to nonnaturalistic moral realists, the reason relation is not itself a relation that we can understand in terms amenable to science, but is rather a primitive *normative* relation.

Modern nonnaturalistic moral realists thus have a ready-made solution to the second and third problems facing Moore. Why cannot goodness float free of the natural properties with which it is coinstantiated? Since goodness is just the property of having some natural feature that provides us with a reason to desire that that very natural feature is instantiated, it follows that when those natural properties are not instantiated, then neither is goodness. And why should we care about the distribution of goodness in the world? We should care because the distribution of goodness in the world just amounts to the distribution of the natural properties in the world that we have a reason to desire be instantiated.

What about the first problem? The first problem, you will recall, is that nonnaturalistic moral realists have to explain how we can come by knowledge of nonnatural features, given that our paradigms of knowledge acquisition are observation and inference. In the case of the reason relation, this problem is especially acute, as the reason relation holds not just between facts and desires, but also, and paradigmatically, between facts and beliefs. In being committed to the irreducibility of the reason relation, modern nonnaturalistic moral realists thus commit themselves to the irreducibility of this relation *even in the case of belief*. But much of epistemology seems to be just an elaborate attempt to explain what this relation is, in the case of belief, in naturalistic terms. So either modern nonnaturalistic moral realists have to suppose that epistemology is a hopeless task, or they have to revise their view.

Of course, it must be admitted that epistemologists do disagree among themselves about what the reason relation is in the case of belief. Some, like Hume, think that we should explain what it is for a fact to provide a reason for believing something in terms of an entailment relation that holds between that fact, on the one hand, and the truth of the thing believed, on the other; others think that we should explain what the reason relation is in terms of raising probabilities; yet others think the key concept is the concept of evidence; and so we could go on. Moreover, these concepts might in turn be further explained in more formal terms – for example, in terms of the probability calculus, or a theory of statistical reasoning – and these theories might themselves be the subject of disagreement. But beneath all of this disagreement lies a general consensus that some such reductive story can be told. The fact that certain facts provide reasons for believing things and others do not is not something that eludes understanding in scientific terms.

This is bad news for the modern nonnaturalistic moral realists. They must either reduce the concept of a reason for desiring itself to the concept of a reason for believing, thereby giving up their nonnaturalism, or else they must admit that the central concept on which their theory is built, the concept of an irreducible reason relation that holds between a fact and a desire, itself bears no relation at all to the reason relation that holds between a fact and a belief. In the end, then, it seems that modern nonnaturalistic moral realism fares no better than Moore's original version.

The Open Question Argument, Nihilism and Expressivism

Nonnaturalistic moral realism seemed to be forced upon us by reflection on the Open Question Argument. The move from the Open Question Argument to nonnaturalism must therefore be flawed. However, there is no great consensus about where, precisely, the flaw in the reasoning lies.

One possibility is that though the argument succeeds in establishing the conceptual truth that we *conceive* of moral features as nonnaturalistic, and hence succeeds in showing that our beliefs about which acts are right and wrong are one and all beliefs about the nonnatural features possessed by acts, Moore went wrong in supposing that any such features are instantiated. Viewed in this light, the problems with nonnaturalism just go to show that such nonnatural features are nowhere instantiated in actuality. No acts have such nonnatural features, and hence all our moral beliefs are false. Accordingly, on this way of thinking about it, the proper conclusion of the Open Question Argument is not a form of moral realism, but rather nihilism. The problem with this, however, is that it concedes the intelligibility of Moorean nonnatural properties, whereas the problems just described make it look like the very idea of a nonnatural property is not really intelligible after all.

Another, and more popular, suggestion has been to suppose that the Open Question Argument constitutes a reductio of the very idea that there are moral features (Hare 1952; Blackburn 1993). According to this suggestion, the reason we cannot come by knowledge of the moral features of acts via an inference from knowledge of their naturalistic features is because that would require that there are two distinct ways the world could be – a naturalistic way and a moral way – which stand in a certain logical relation to each other. But there are not two ways the world could be: there is only one way, a naturalistic way. The upshot, according to this suggestion, is that the claim that the world is a certain way morally is not true, strictly speaking. The role of a moral claim is not to represent the world as being a certain way, and hence is not to give the content of any belief, but is rather to express desires or aversions. Thus, according to this way of thinking, the Open Question Argument constitutes a second, and many think much more decisive, line of argument for expressivism.

The problem with this way of thinking about Moore's Open Question Argument, however, is that it assumes that expressivism itself somehow manages to escape the clutches of the Open Question Argument (Smith 1998). In fact, however, expressivism is vulnerable to a version of the argument. This is because though expressivism sets itself against the view that (say) "Torturing babies is wrong" is analytically equivalent to some naturalistic claim about the way the world is – for that would assume something expressivism takes to be false, namely that wrongness is a feature of acts – it does so by insisting that a sentence like "Michael judges that torturing babies is wrong" *is* analytically equivalent to some naturalistic claim. Specifically, expressivism takes this sentence to be analytically equivalent to "Michael expresses his aversion to torturing babies." But now it seems that we can run the following version of the Open Question Argument against expressivism.

If "Michael judges that torturing babies is wrong" and "Michael expresses his aversion to torturing babies" were analytically equivalent, then the question "Michael expresses his aversion to torturing babies, but does he judge that torturing babies is wrong?" would have to be one whose answer is immediately obvious to anyone who understands the meanings of the words used. However, as a moment's reflection reveals, the questions clearly are not equivalent. We can, without self-contradiction, agree that Michael expresses his aversion to torturing babies, but deny that he thereby judges it to be wrong. But, if this is right, then it follows that the question "Michael expresses his aversion to torturing babies, but does he judge it to be wrong?" is not a closed question – one whose answer is immediately obvious to anyone who understands the meanings of the words – but is rather an open question – one whose answer is open to reasoned argument. "Michael expresses his aversion to torturing babies" and "Michael judges that torturing babies is wrong" are thus not analytically equivalent. They do not pick out the same feature of the world.

Moreover, as with the earlier application of the Open Question Argument, it does not seem to matter which of the various natural features of Michael we consider. It would always be open to reasoned argument whether, when Michael expresses any of the various natural features we might care to consider – various complexes of desire, second-order desires, or whatever – he is thereby judging that torturing babies is wrong. If sound, the Open Question Argument therefore seems to show that Michael's judging it wrong to torture babies is not analytically equivalent to any natural feature of Michael either. If the Open Question Argument refutes naturalistic moral realism, it refutes expressivism as well.

But now we have surely proved too much. After all, we said at the outset that the only options available are nihilism, expressivism, or some form of moral realism – that is, either naturalistic or nonnaturalistic moral realism. Yet what we have just seen is that, if sound, the Open Question Argument, together with ancillary premises, rules out *all* these options. That surely cannot be. The only conclusion to draw is therefore that, properly understood, the Open Question Argument is *unsound*. But wherein lies the mistake in the argument?

The Naturalistic Moral Realist's First Response to the Open Question Argument

Many contemporary naturalistic moral realists argue that the flaw lies in the assumption that it somehow follows from the fact, conceding it to be a fact, that "rightness" and "the property of maximizing utility" are not analytically or a priori equivalent, that these terms pick out different features. That this is a flaw is, they insist, evident from examples with which we are familiar in empirical science (Brink 1989; Darwall, Gibbard, and Railton 1992).

For example, "Water" is not analytically or a priori equivalent to "H_2O," but empirical science teaches us that water is just H_2O. "Redness" is not analytically or a priori equivalent to "surface reflectance property α," but empirical science teaches us that redness is just a certain surface reflectance property that we will, for convenience, call "α." So, in this particular case, they argue that the Open Question Argument assumes, wrongly, that if rightness and the property of maximizing utility were one and the same property, then it would have to be an a priori truth, one discovered by reflection on the meanings of the words "rightness" and "the property of maximizing utility." But it thereby overlooks the possibility that it may be an a posteriori truth, one discovered through observation and inference. Naturalistic moral realists who offer this reply to the Open Question Argument therefore face a challenge. They must show how it could be an a posteriori truth that these terms pick out the same feature. Unfortunately, however, those who face up to this challenge find that they simply run into the Open Question Argument all over again.

According to many naturalistic moral realists, for example, the reason it is an a posteriori truth that rightness is the property of maximizing utility is that we invoke obligatoriness, rightness, and wrongness in order to explain various empirical phenomena, and then we discover, a posteriori, that the maximization of utility occupies the relevant explanatory role. For example, they argue that since, contingently, right actions have certain effects – they are causally responsible for a tendency toward social stability, for example – so it follows that we can fix the reference of the term "right" via the description "the property of acts, whatever it is, that is causally responsible for their tendency toward social stability." Equipped with this reference fixing description, we can then investigate acts with this effect in order to find out which feature explains this tendency. If, say, we discover that the feature that is causally responsible is the maximization of utility, then we can conclude that rightness is the property of maximizing utility. Our conclusion will then be a posteriori, not a priori.

The answer is supposed to be straightforward because the explanation involved has the same structure as those we give in other less controversial cases. Since, contingently, red objects have certain effects – they cause those objects to look red to normal perceivers under standard conditions – so it follows that we can fix the reference of "redness" via the description "the property of objects, whatever

it is, that causes them to look red to normal perceivers under standard conditions." Equipped with this reference-fixing description we can then investigate the acts which have this effect in order to find out which feature explains this tendency. If, say, we discover that the feature that is causally responsible is surface reflectance property α, then we can conclude that redness is surface reflectance property α.

Unfortunately, however, this reply to the Open Question Argument is inadequate, and the reason why is perhaps already evident (Jackson 1997). Consider again the case of colors. True enough, we will not find a justification for thinking that redness is surface reflectance property α merely by reflecting on the meanings of the words "surface reflectance property α" and "redness." The fact that redness is surface reflectance property α is clearly something we discover a posteriori through empirical investigation. But in explaining how we come to make this discovery a posteriori, it is clear that we do in fact appeal to an a priori truth about redness. For we simply assumed that we can fix the reference of "redness" via the description "the property of objects, whatever it is, that causes them to look red to normal perceivers under standard conditions." But what sort of justification can we give for this claim? It clearly is not supposed to be yet another a posteriori truth. Rather it is supposed to be an a priori truth – one which is either stipulated in the act of reference fixing itself or else discovered by reflection on the everyday meaning of the word "red." Either way, it is because we accept this claim a priori that we can move straight from the discovery that surface reflectance property α is the property that causes objects to look red to normal perceivers under standard conditions to the conclusion that surface reflectance property α is redness.

By analogy, then, even though it may well be an a posteriori truth that rightness is the property of maximizing utility, in the very argument we gave in support of this claim it is clear that we in fact appealed to another truth, but this time one which is supposed to be known a priori, about the relation between rightness and certain natural properties. For we simply assumed that we could fix the reference of "rightness" via the description, "the property of acts, whatever it is, that is causally responsible for their tendency toward social stability." But what sort of justification can be given for this claim? It clearly is not supposed to be another a posteriori truth. Rather, it is supposed to be an a priori truth – one which is either stipulated in the act of reference fixing or else discovered by reflection on the everyday meaning of the word "right." Either way, it is precisely because we accept this claim a priori that we can move straight from the discovery that the property of maximizing utility is the property acts possess when they tend toward social stability to the conclusion that the property of maximizing utility is rightness.

This is an extremely important point – one which is quite devastating to those naturalistic moral realists who think they can reply to Moore's Open Question Argument by insisting that, even though the terms "rightness" and "the property of maximizing utility" are not analytically or a priori equivalent, these terms nonetheless pick out the same feature of acts. For what they fail to remember is that Moore's Open Question Argument is supposed to refute *all* claims to the effect that "rightness" is analytically or a priori equivalent to a term ascribing natural

features to acts, no matter which natural features are in question. If sound, it thus even refutes the claim that it is a priori that rightness is the property acts possess when they tend toward social stability. The alleged refutation goes like this: we can agree that an act has the property which is causally responsible for the tendency of acts toward social stability and yet, apparently without self-contradiction, deny that it is right, for it is an open question whether such acts are right – a matter for reasoned argument. "Rightness," we should thus conclude, cannot be a priori equivalent to "the property acts possess when they tend toward social stability."

This teaches us a valuable lesson. Naturalistic moral realists have no alternative but to face head-on the claim that we can, via the Open Question Argument, refute the claim that there is some naturalistic analytic or a priori equivalent for "rightness."

The Naturalistic Moral Realist's Second Response to the Open Question Argument

Moore claims to show that there is no naturalistic claim that is analytically or a priori equivalent to any moral claim by pointing out that it is always an open question whether an act with whichever naturalistic features we care to choose has some moral feature. It is always open to reasoned argument. In order to see where this argument goes wrong, we need to think more generally about the project of conceptual analysis (Smith 1994: ch. 2; Jackson 1997).

When we try to analyze a concept, what are we trying to do? The answer is roughly this. There are all sorts of constraints on the way we use various words. Consider color words as an example. It is a constraint on the proper use of color words that we use them to pick out properties that cause us to have certain visual experiences; a constraint that we use them to pick out features that are more reliably detected in daylight than in the dark; a constraint that people's use of them is especially likely to be defective if there is something wrong with their eyes; and so on and so forth. When we try to come up with an analytic equivalent of "x is red," our task is to come up with something that captures this complex set of constraints; that is, to come up with an account of what "redness" means that entails them. When we say that "redness" and "the property of objects that causes them to look red to normal perceivers under standard conditions" are analytically or a priori equivalent this is what we have in mind.

If this is right, however, then the success or failure of an analysis is to be judged accordingly. It is not to be judged by the obviousness of the analysis; nor is it to be judged by whether the analysis is open to reasoned argument. Indeed, if what we have just said is right then it will of course be open to reasoned argument whether or not an analysis is successful because it will be open to reasoned argument what the complex set of constraints on the use of the word being analyzed is and whether or not this complex set is entailed by the proposed analysis.

If an account of the project of conceptual analysis along these lines is right, however, then Moore's argument evidently fails altogether to refute the claim that "rightness" has a naturalistic analytic or a priori equivalent. Consider, for example, the claim that "rightness" is analytically equivalent to "the property of acts that is causally responsible for their tendency toward social stability." It is irrelevant whether or not it is obvious that this is so; irrelevant whether it is open to reasoned argument. The only relevant question is whether, on reflection, we think that this analysis entails the complex set of constraints on the way in which we use the word "right." If it does, then it is analytically equivalent, notwithstanding the fact that it is not obvious.

In many ways this brings us to where we are today in the moral realism debate. The Open Question Argument is not sound, but it does make clear the enormous task that lies before naturalistic moral realists. To repeat, naturalistic moral realists must give a naturalistic account of the contents of moral beliefs; an account of the naturalistic feature that they take to be identical with various moral features. But what the Open Question Argument brings out is that, in doing so, they must find naturalistic features that are analytically – or, anyway, a priori – equivalent to those moral features. It need not be obvious that the naturalistic features and moral features are analytically equivalent, of course. It may be open to reasoned argument. But, at the end of the day, it must be demonstrable, on the basis of reflection on the ways in which we use moral words, that the naturalistic features they identify are one and the same as moral features.

Externalist Naturalistic Moral Realism

The naturalistic theories that have been claimed to fit this bill fall into two quite distinct categories. The first are versions of externalist naturalistic moral realism (hereafter "externalist realism" for short) (Sturgeon 1985; Railton 1986; Brink 1989). This is the view that though we can, by reflecting on the ways in which we use moral words, find a naturalistic equivalent for the term "rightness," the naturalistic equivalent we come up with will leave it completely open whether someone who believes an act to be right will desire to perform that act, or be indifferent to performing it, or be averse to performing it. The sort of theory described earlier which claims that rightness fills a distinctive explanatory role, the role of underwriting a tendency toward social stability, is an example of such a theory. Externalist realists thus face a dual task.

On the one hand, they must come up with an explanation of why, when we reflect on the way in which we use moral words, we should conclude that rightness has the naturalistic equivalent they posit. For example, if we consider again the theory which holds that rightness is the property of acts, whatever it is, that is causally responsible for their tendency toward social stability, externalist realists

must tell us what it is about the way in which we use moral words that is supposed to make this particular claim seem credible. The vast literature on moral explanations is perhaps best seen as addressing this issue (Harman 1977; Sturgeon 1985; Railton 1986; Boyd 1988). As I understand it, the claim externalist realists make in that literature is that "rightness" is a term whose meaning is fixed by a causal explanatory theory which assigns rightness a certain characteristic explanatory role. "Rightness" is thus, in a sense, much like the term "electron." Both terms serve to pick out a feature in virtue of the characteristic causal role that that feature occupies.

On the other hand, however, externalist realists must also try to explain away the fact that so many people have been inclined to think that our use of moral words is subject to the internalist constraint. If what we believe, in believing an act to be right, is (say) that the act has the feature that is causally responsible for a tendency toward social stability, then why have so many people been inclined to think that possession of this belief requires a desire to perform such an act, at least other things being equal, that is, absent depression, weakness of will, and the like? It is surely an entirely contingent matter whether someone with such a belief will desire to perform such an act, whether they are depressed and weak of will or not. So how did so many philosophers get it wrong for so long? For many, the very fact that externalist realism is incapable of capturing the internalist constraint is a decisive reason to reject the theory. But though I am inclined to agree with this objection, I do not want to rest the case against externalist realism wholly on it.

Suppose we grant the idea that "rightness" picks out a property in virtue of its explanatory role. Still, must not the explanatory role in question be one that somehow guarantees the possibility of giving a justification for acting in the way that is deemed to be right (Sayre-McCord 1988b)? After all, by all accounts, the fact that an act is right implies that there is at least some justification for performing it. Someone who says "Though it would be right to act in that way, there is no justification at all for doing it" misuses the word "right." Yet the most remarkable feature of externalist realism is that it makes this connection altogether mysterious. Focus again on the version of externalist realism developed earlier. The most remarkable feature of the suggestion that rightness is that property, whatever it is, possessed by acts that tend toward social stability, must surely be that an act may conduce to social stability but be one that there is no justification *at all* for anyone's performing. Explanatory role and justificatory potential just seem to be quite different things.

At the end of the day, then, the really difficult task facing externalist realism is thus to come up with an account of the explanatory role of rightness which makes that role connect in some constitutive way with the possibility of giving a justification. Until externalist realists come up with such an account, their theory will look like it fails to capture one of the most important constraints on the way in which we use moral words.

Internalist Naturalistic Moral Realism

This brings us to what seems to me to be the most plausible version of moral realism. Internalist naturalistic moral realists (hereafter "internalist realists" for short) agree with externalist realists that we can characterize rightness in terms of its distinctive explanatory role, but they hold that the explanatory role characteristic of rightness is, broadly speaking, that of eliciting desire under certain idealized conditions of reflection. Consider a specific version of the theory, by way of illustration (Smith 1994).

According to this version of the theory, obligatoriness is that feature, whatever it is, that we would desire our acts to possess if our desires formed a set that is maximally informed, coherent, and unified. The internalist realist's claim is that this analysis of obligatoriness finds support in the way in which we use moral words. It is not difficult to see what reasons they might give. After all, as we have just seen, when we say that acting in a certain way in certain circumstances is obligatory, we thereby imply that there is some justification for our acting in that way in those circumstances. But facts about what there is a justification for doing, in various circumstances, are in turn plausibly thought to be facts about what we would advise ourselves to do if we were better placed to give ourselves advice; that is, more precisely, they are plausibly thought to be facts about what we would desire ourselves to do in those circumstances if our desires were immune to rational criticism. That is just what the theory says.

Of course, it might be thought that there are other ways of thinking about justification. But internalist realists argue that this particular analysis of the notion is amply supported by various other ways in which we use moral words. For example, it is agreed on nearly all sides that moral knowledge is a relatively a priori matter, at least in the following sense: if you equip people with a full description of the circumstances in which someone acts, then they can figure out whether the person acted obligatorily, rightly, or wrongly by just thinking about the case at hand. Someone who claimed that it would be impossible to figure out what is obligatory by just thinking about the circumstances of action would be misusing the word "obligatory." Internalist realists argue that this is well explained by the analysis just offered. It is because we can subject our desires about what is to be done in various circumstances to critical evaluation by just reflecting on our desires that moral knowledge seems to be such a relatively a priori matter.

Internalist realists also claim that the fact that there is connection between what it is obligatory to do and what there is a justification for doing in turn explains the internalist constraint on the use of moral words. Suppose you believe that a certain act available to you is one that you would desire yourself to perform if you had a set of desires that was maximally informed and coherent and unified. You are then arguably under some rational pressure to have a corresponding desire. After all, desiring to act in the way you believe you would want yourself to act if you had a maximally informed and coherent and unified desire set coheres better

with, or fits better with, or makes more sense in the context of, that belief, than would being either averse to or indifferent to acting in that way. The coherence of your psychology thus seems to demand the desire of you.

Internalist realists insist that it should therefore come as no surprise at all that those who believe that acting in a certain way would be right will desire to act in that way, at least absent the effects of depression, weakness of will, and the like. Indeed, they argue that their analysis serves to reveal the essential nature of depression, weakness of will and the like. As psychological conditions that can undermine the connection between moral belief and desire, depression, weakness of will, and the like share a common feature: they are all conditions with the inbuilt potential to create psychological incoherence. No surprise then that, absent the conditions that make for that sort of incoherence, people will desire to act in the ways that they believe they would want themselves to act if they had a maximally informed and coherent and unified desire set.

Should an Internalist Naturalistic Moral Realist Be a Relativist?

For these and other reasons, internalist realists think that their own theory is therefore a vast improvement on externalist realism. There is, however, an important ambiguity in the internalist realist's theory that still needs to be addressed. This touches on the issue of relativism.

Obligatoriness is supposed to be that feature, whatever it is, that we would desire our acts to possess if our desires formed a set that is maximally informed, coherent, and unified. But is the idea supposed to be that the "we" referred to in the analysis includes all rational creatures? In other words, is the idea that we would all converge in the desires we would have under idealized conditions of reflection? Or does the "we" include only some subset of the rational creatures? Does it include, say, me and those who desire things similar to the things that I actually desire? In other words, are contingent and rationally optional culturally induced differences in our actual desires supposed to make convergence in the desires we would have under conditions of idealized reflection impossible?

If the latter then the theory is relativistic (Harman 1975, 1985). According to relativistic internalist naturalistic moral realism, when we say that actions of a certain sort are obligatory what we are really saying is a subset of rational creatures – those who have desires like our own – are such that they would desire that we act in that way if they had desires that formed a maximally informed, coherent, and unified set. However, we thereby allow that other perfectly rational creatures may differ from us. This need not force us to think that their acting in that way would not be obligatory as well. If we believe that we would desire them too to act in that way under idealized conditions of reflection then of course we will believe that acting in that way would be obligatory for them too. The crucial point is simply that their having corresponding desires as part of their idealized desire

set is no part of what makes our claim that it is obligatory for them so to act true. On the alternative analysis, by contrast – that is, according to nonrelativistic internalist naturalistic moral realism – their possession of such desires too is required for the truth of our claim (Smith 1994).

But of these two versions of the internalist realist's theory it seems to me that the relativistic version is manifestly implausible as conceptual analysis. How could, whether or not an act is obligatory, right, or wrong and hence is justified or unjustified, the paradigm of a nonarbitrary fact about an act be grounded in something so arbitrary as whether or not someone happens to have certain contingent and rationally optional culturally induced desires? The very idea seems to involve a contradiction. Yet this is the conclusion to which the internalist realist who buys into relativism is committed.

The nonrelativistic version of the theory, by contrast, holds that such facts are grounded in something that is itself appropriately nonarbitrary. Acts are obligatory, right, or wrong depending on whether, notwithstanding any contingent and rationally optional culturally induced differences in our actual desires, we would all desire or be averse to the performance of such acts if we had a set of desires that was maximally informed, coherent, and unified. Underlying this form of internalist realism is thus a resonant picture of ourselves and our relations to other people. At the very deepest level – that is, in that idealized possible world in which we all have a set of desires that is maximally informed, coherent, and unified – we share common aims simply in virtue of our nature as rational beings. No one is beyond the pale: not, at any rate, if they remain susceptible to rational argument. Even the most wretched may be reachable.

Notwithstanding the resonance, this picture may of course still be all mere illusion. The truth of the nonrelativistic version of internalist realism depends on more than mere conceptual analysis. It depends, as well, on the substantive fact that *there is* a set of desires that we would all converge upon if we had a set of desires that was maximally informed, coherent, and unified. Even if the conceptual analysis is impeccable, absent the power of rational argument – that is, absent the power of information, together with considerations of coherence and unity – to elicit common desires in us, the nonrelativistic version of internalist realism entails that there are no moral facts at all. Unsurprisingly, this means that the truth of the nonrelativistic version of internalist realism requires us to carry out a version of the traditional rationalist project in moral philosophy (Smith 2011). We have to show that mere reflection on what it is to have a maximally informed and rational psychology allows us to see that certain desires are a part and parcel of every such psychology, where these desires are in turn those that fix the content of morality. But the traditional rationalist project is, of course, fraught with difficulty.

The proper conclusion to draw is thus that even the very best version of moral realism is sub judice, something about which we will be convinced only to the extent that we are confident that the arguments we give ourselves for desiring as we do are arguments that should convince the arbitrary rational person to desire likewise. And, of course, experience teaches that that kind of confidence is difficult

to maintain. The unfortunate tendency of the media to portray people from other cultures as radically different from each other, as though they do not even share a common tendency to believe and desire on the basis of reflection as opposed to superstition, let alone a common tendency to desire alike after reflecting, doubtless plays a significant role. Even convinced nonrelativistic internalist naturalistic moral realists will therefore continue to feel the pull of nihilism in their more pessimistic moments.

References

Blackburn, S. (1984) *Spreading the Word*, Oxford: Oxford University Press.

Blackburn, S. (1993) "Circles, Finks, Smells and Biconditionals," *Philosophical Perspectives* 7: 259–79.

Blackburn, S. (1994) *Essays in Quasi-realism*, New York: Oxford University Press.

Boyd, R. (1988) "How to Be a Moral Realist," in *Essays on Moral Realism*, ed. G. Sayre-McCord, Ithaca, NY: Cornell University Press, pp. 181–228.

Brink, D. (1989) *Moral Realism and the Foundations of Ethics*, Cambridge: Cambridge University Press.

Carroll, L. (1872/1998) *Alice's Adventures in Wonderland and Through the Looking Glass*, ed. R.L. Green, Oxford: Oxford Paperbacks.

Darwall, S., Gibbard, A., and Railton, P. (1992) "Toward *Fin de Ciecle* Ethics: Some Trends," *Philosophical Review* 101 (1): 115–89.

Dreier, James (2004) "Meta-ethics and the Problem of Creeping Minimalism," *Philosophical Perspectives* 18 (1): 23–44.

Gibbard, A. (1990) *Wise Choices, Apt Feelings*, Oxford: Clarendon Press.

Hare, R.M. (1952) *The Language of Morals*, Oxford: Oxford University Press.

Harman, G. (1975) "Moral Relativism Defended," *Philosophical Review* 84 (1): 3–22.

Harman, G. (1977) *The Nature of Morality*, Oxford: Oxford University Press.

Harman, G. (1985) "Is There a Single True Morality?" in *Morality, Reason and Truth*, eds. D. Copp and D. Zimmerman, Totowa, NJ: Rowman & Allanheld, pp. 27–48.

Horwich, P. (1990) *Truth*, Oxford: Blackwell.

Horwich, P. (1993) "Gibbard's Theory of Norms," *Philosophy and Public Affairs* 22 (1): 67–78.

Jackson, F. (1997) *From Metaphysics to Ethics*, Oxford: Oxford University Press.

Jackson, F., Oppy, G., and Smith, M. (1994) "Minimalism and Truth-Aptness," *Mind* 103 (411): 287–302.

Joyce, R. (2001) *The Myth of Morality*, Cambridge: Cambridge University Press.

Mackie, J.L. (1977) *Ethics: Inventing Right and Wrong*, Harmondsworth, UK: Penguin.

Moore, G.E. (1903) *Principia Ethica*, Cambridge: Cambridge University Press.

Parfit, Derek (2011) *On What Matters*. 2 vols. Oxford: Oxford University Press.

Railton, P. (1986) "Moral Realism," *Philosophical Review* 95 (2): 163–207.

Railton, P. (1993a) "What the Noncognitivist Helps Us to See the Naturalist Must Help Us to Explain," in *Reality, Representation and Projection*, eds. J. Haldane and C. Wright, Oxford: Oxford University Press, pp. 279–300.

Railton, P. (1993b) "Reply to David Wiggins," in *Reality, Representation and Projection*, eds. J. Haldane and C. Wright, Oxford: Oxford University Press, pp. 315–28.
Sayre-McCord, G., ed. (1988a) "The Many Moral Realisms," in *Essays on Moral Realism*, Ithaca, NY: Cornell University Press, pp. 1–23.
Sayre-McCord, G., ed. (1988b) "Moral Theory and Explanatory Impotence," in *Essays on Moral Realism*, Ithaca, NY: Cornell University Press, pp. 256–81.
Scanlon, Thomas M. (1998) *What We Owe to Each Other*, Cambridge, MA: Harvard University Press.
Smith, M. (1994) *The Moral Problem*, Oxford: Blackwell.
Smith, M. (1998) "Ethics and the A Priori: A Modern Parable," *Philosophical Studies* 92 (1–2): 149–74.
Smith, M. (2011) "Deontological Moral Requirements and Non-welfarist Agent-Relative Values," *Ratio* 24 (4): 351–63.
Sturgeon, N. (1985) "Moral Explanations," in *Morality, Reason and Truth*, eds. D. Copp and D. Zimmerman, Totowa, NJ: Rowman & Allanheld, pp. 49–78.
Wright, C. (1992) *Truth and Objectivity*, Cambridge, MA: Harvard University Press.

Chapter 2

Relativism

Simon Blackburn

Relativism in ethical theory is the doctrine that ethical truth is somehow relative to a background body of doctrine, or theory, or form of life or "whirl of organism." It is an expression of the idea that there is no one true body of doctrine in ethics. There are different views, and some are true for some people, while others are true for others.

This has affinities with a practical stance that encourages toleration of different societies or different approaches to practical living. But such toleration could in principle coincide with an absolutist theory of ethics, according to which there is just one correct body of ethical doctrine. For that one correct body of doctrine could, in principle, include the view that it is permissible or even obligatory to tolerate people who do things differently. Nevertheless, the practical attitude of fairly universal toleration is often felt to be a *consequence* of the theoretical stance that there is no one truth. That is, once the theorist takes the view that there are pluralities of ethical truths, each relative to the different positions of people, it becomes quite natural to draw the conclusion that toleration is the only warranted stance. For if "they have their truth" and we have ours, it would seem at best a brute exercise of power to coerce them into our ways, or ostracize them, or go to war with them for doing it differently. Thus the postmodernist Richard Rorty took himself to have dismantled any conception of absolute ethical truth, and indeed drew the conclusion that only a light, ironic, aesthetic stance to practical problems was justified (Rorty 1989). Rorty was not unique in this. Many people, and many ethical theorists, believe that without some "robust" or "objective" conception of moral truth, our *right* to hold judgments with a sufficient degree of conviction evaporates. If we want to oppose cruelty, or defend free speech, or outlaw child sex, we need the conviction that it is not "just us," voicing a contingent or

The Blackwell Guide to Ethical Theory, Second Edition. Edited by Hugh LaFollette and Ingmar Persson.

accidental aspect of how we feel. We want to hold that truth is on our side: absolute truth, even God's truth. This is why relativism is usually seen as a disturbing challenge to moral authority: a challenge that it is the business of a proper moral theory to answer.

In this essay I shall compare a number of approaches to ethical theory, in terms of how they react to relativism. I shall start with the theory I myself defend, which is a kind of "projectivism" or "expressivism." I shall argue in the next section, "Expressivism and Relativism," that, perhaps surprisingly, this metaphysically undemanding view of ethics in fact puts us in a very strong position to make the right reaction to relativism. I then say how certain other approaches, while superficially better able to defend themselves against relativism, in fact open up disturbing relativistic options. While ostensibly protecting our authority, they in fact undermine it.

Expressivism holds that the key to ethics lies in the practical stances that we need to take up, to express to each other, and to discuss and negotiate. As the expressivist tradition has always emphasized, ethics is at bottom about practice. It is about choices and actions. Ethics concerns what to do, and what not to do; who to admire and who to avoid; where to draw lines, and where not to; which feelings to cultivate, and which ones to repress. Our ethic is shown in our cluster of dispositions to encourage and to discourage various choices, characters, and feelings. A sincere moral opinion is the expression of one of these dispositions. For this reason ethics can, fundamentally, be expressed in terms of prescriptions, and in some systems, like that of the Old Testament, this is indeed how it is given: thou shalt do this, and not do that. But prescriptions only get us part of the way. Attitudes need comparison and ranking as well as simple expression. "This is better than that" can get expression as "admire this more than that," but the replacement would be strained. And we have an ethical language that goes beyond simple imperatival forms for good and sufficient reason, namely to bring to the business of systematizing and reasoning about attitude the elegant framework of propositional logic.

The reason for this is that practice is so important. Hence our moral and evaluative dispositions need discussion. They need to be queried, and sometimes qualified and rejected and replaced. These queries can take the form of asking whether a particular opinion is *true* or *right*. But the appearance should not mislead us. As Wittgenstein often reminds us, *p* is true means that *p*.[1] Asking whether a moral judgment is true or right is no more than asking whether to accept it. And asking that is asking which attitude or policy or stance to endorse.

All this should be platitudinous. But in fact in many people's minds it rings all kinds of alarm bells. It sounds to them to be an invitation to some alarming downgrading of ethics. It seems to undermine the "absolute" or binding or authoritative character that we associate with moral and ethical imperatives.[2] Some people who share the basic orientation have encouraged these anxieties. Bernard Williams believed that ethics cannot be what it seems (Williams 1985). And famously, John Mackie, in his "error theory," supposed that there are elements in

our ethical practice that could only be justified if something more were true: some kind of "objectivity" or "authority" or "to-be-doneness" built into the frame of the world. But this something more is not true, so first-order ethical practice is founded on a mistake (Mackie 1977).

The expressivist takes these views to be natural, but also to be rejected: first-order ethical practice embodies no mistake at all. Particular ethical views, of course, may be mistaken. But the categories of ethics and the states of mind of those who find they need them and express themselves in terms of them are quite in order. Or rather, if they are not (as some people think that the concept of a "right," for example, is of doubtful use) then it is an ethical problem, not one of logic or metaphysics or the theory of what ethics is. One sign of this is that Mackie himself never showed what a practice of expressing and comparing and encouraging and discouraging practical stances would look like if it were *free* of the mistake he alleges. One might suppose that it would come to look exactly like our own practice. And this shows that there is no mistake embedded in the very structure of our reasonings.

Expressivism and Relativism

There is no problem of relativism for the expressivist, because there is no problem of moral truth. Since moral opinion is not primarily in the business of *representing* the world, but of assessing choices and actions and attitudes in the world, to wonder which attitude is right is to wonder which attitude to adopt or endorse. Suppose, then, to take a real-life example, that I adopt and express an attitude: say, that women should be educated. Suppose I meet a member of the Afghan Taliban, who holds the reverse. This may certainly pose me a practical problem: in fact at least two practical problems. First, I would like to be able to change his attitude. And second, even if I cannot do that, I would like to be able to stop him from implementing it. But I may not know how to go about either of these things, and there lie my practical difficulties.

The relativist will get up at this point and say, "Well, it is true for the Taliban that women should not be educated." But what can that mean? Surely it is just a bad way of saying that the Taliban *hold* that women should not be educated, which we already knew. It is not a way of putting that opinion in a favorable light. "True for them" sounds a qualified kind of truth, like "It is true in Greenland that it freezes all winter," as opposed to "It is not true in Carolina that it freezes all winter." But seeing the view of the member of the Taliban as true in *any* kind of way must involve putting it into a *favorable* light. And if anyone *wants* to put that opinion in a favorable light, that person has some very hard work to do, and we can promise in advance that he or she will fail. Why? Because nothing worth respecting speaks in favor of the view. There is no favorable light in which it can be put; it is a view that can only appear attractive when the light is very dim.

If you think otherwise, I am against you and I will express this by saying that you are wrong. I reject your opinions, and this I voice by word and deed.

The relativist will try again, saying, "Well, it is *merely* your attitude against his." Part of this is right: it is indeed my attitude against his. That is what ethical conflict is. The part that is wrong is the "merely." What is "mere" about a conflict of attitude? The world's worst conflicts are those of policy, choice, and practice. They are the most important conflicts there are. By comparison, mere conflicts of opinion can fade into insignificance. It need not matter at all to me that you hold that the distance of the moon is half a million miles, although I hold that it is nearer a quarter of a million. The difference need not translate into actions that bother me. But if your attitude to me is contempt or disgust, that matters a great deal, and if your attitude is that women should not be educated, my humanity rebels against you.

No doubt this mistakes the intended import of the word "mere." It is not that conflict of attitude is unimportant. The relativist will say, "It is your attitude against his and *neither of you can show that the other is wrong*." The conflict is "merely" a conflict of attitude in the sense that there is no proof procedure. This, I should say, contains a grain of truth, although only a very small grain. For, after all, it is strictly false. I can show that the Taliban member is wrong by the simplest means: any educated female is a perfectly good illustration of his error. My wife shows how wrong he is, and so do millions of other women. Probably the complaint is that I cannot *show the member of Taliban himself* that he is wrong for, after all, he is blind to the illustration or takes it the wrong way. But even this is not axiomatic. It may be possible to show him himself that he is wrong. We may be able to increase his experience of women, to undermine the delusive authorities on which he relies, to enlarge his sympathies, and so on. It is unlikely to be a *quick* process, but whoever thought that it should be? It is also probably a process that is more likely to be successful if attempted by someone nearer to the Taliban member's frame of mind: a more liberal Muslim, for example.

The only grain of truth in the remark is that there is no algorithm for success. There is no proof procedure, nor any empirical process of working on the member of the Taliban, that is *guaranteed* in advance to bring him to my opinion. But that is just how it is. It is always contingent, and sometimes chancy, whether we can move a dissident toward concurrence with our own sympathies and attitudes. If that worries anyone, they would do well to reflect that the same is true in empirical and even mathematical or logical cases. The problem with the Taliban member is that he is blind to what illustrates his error, and that may be true of people in these other cases. I can show that daffodils are yellow, but I cannot necessarily show to some particular dissident that they are yellow, if he refuses to look, or looks but is blind to colors. I can show that contradictions are false, but I cannot necessarily show it to some enthusiast who holds in advance that all logic is a patriarchal plot of which I am a part.

One formula that has attracted some admiration in recent discussions of relativism is that on such an issue we can simply say, "There is nothing else to think"

(Wiggins 1990/1991). This may just be a flowery way of nailing our colors to the mast, but otherwise it strikes me as overdoing it. I think it is unfair to the member of the Taliban to deny that he indeed thinks something else. It dehumanizes him, making him into some kind of mad dog, unable to think about choice, action, and ways of life at all. Whereas in fact he voices a genuine attitude, and a practical policy, for which he has his delusive reasons. That is why we are in conflict. Otherwise he would represent nothing but a kind of obstacle to our own policies; something, in Peter Strawson's words, to be "managed or handled or cured or trained" (Strawson 1968). I have no sympathy with the Taliban member, but I would not myself wish to put him outside the pale of human thought. We cannot conjure away attitudes of separatism and divisiveness, either of class or race or gender, by a kind of legislation that the people who hold them are beyond interpretation.

"Well now," says the relativist as a last resort, "but is not expressivism usually classed as an antirealist view of ethics?" (That is, a view denying that ethical facts or properties have the same kind of status as real facts or real properties.) "And does not this mean that *there are no facts* of an ethical or normative kind? And then are not all attitudes fundamentally on an equal footing?" The answers are: yes, expressivism is usually characterized as antirealist, although that label has its problems; but no, this does not mean that there are no facts of an ethical or normative kind; and finally, even if it did, this would not mean that all attitudes are on all fours.

Expressivism which is tolerant of the forms that ethical thought naturally takes is often called quasi-realism: there is an appearance of realism without any metaphysical cost. It refuses to give ethical facts a typical explanatory role. This is already heralded when we turn our backs on ethical representation. A representation of something as F is typically explained by the fact that it is F; a representation *answers to* what is represented. Expressivism holds that ethical facts do not play this explanatory role. We cannot, except by analogy, talk of ethical perception. If we want a slogan we can say that the way of the world, and that includes the mental world, is independent of the evaluative and the normative. Oughts do not explain is's. Our moral understandings are not explained by independent moral structures, to which we are lucky enough to be sensitive.

Why does this not imply that there are no moral facts? Minimalism shows us why not. I have already given you a moral opinion of mine: women should be educated. Here is another way of putting it: it is true that women should be educated. Here is another: it is a fact that women should be educated. If we like we can go further up this progression, which I call Ramsey's ladder: it is true that it is a fact . . . ; it is really true that it is a fact . . . ; and, soon, but not quite yet, I shall suggest that we can add objectivity to the list.[3]

And why does that not imply that divergent moral opinions are on an equal footing? Well, all the expressivist can hear that as meaning is that they are all *equally good*. And that is just not true. The Taliban member's opinion on the education of women is not as good as mine. In fact, it is diametrically wrong – wrong root

and branch. But notice that this would be so even if we were less minimalist than I have been about facts. Suppose a substantive or robust theory of truth were developed, giving us some notion of correspondence. Suppose it proceeds by isolating some metaphysical category of Facts (note the uppercase). And suppose finally that for the kinds of reason I have outlined, there are no normative or ethical Facts (all these doctrines belong to the *Tractatus Logico-Philosophicus*). This would be a metaphysical result. So it clearly could not imply that all moral opinions are on an equal footing. It could not imply, for instance, that it is permissible to hold that women should not be educated. It could at best imply that in holding this you do not trespass against the uppercase Facts. But that is all right. It is not *that* (or, not simply that) that is wrong with the Taliban view. The main thing that is wrong with the view is that it is inhumane, cruel, arbitrary, and so on. The metaphysics cannot imply that it is all right to be like that!

I have said that there is no problem of relativism, and tried a little to explain why this is so. I shall finish this part of the discussion by entering a very small concessive remark: something that can perhaps serve to salvage a little pride for the relativist.

There are cases like that of the Taliban, but there are also cases where travel broadens the mind. We might start off by thinking that our attitude is the only permissible attitude, or our ways are the only permissible ways, and that all others are wrong. But exposure to other people, or other cultures or times can make us change our minds. They do it differently; yet we cannot condemn them, or find it in our hearts to maintain the superiority of our ways. So we become a degree more tolerant. And this is often exactly as it should be.

I suspect that the relativist generalizes too rapidly from this kind of progression, assuming that because it is as it should be in some cases, it must be so in all cases. Hence simple exposure to alternative opinion should be enough to dissolve any allegiance we hold to our own attitudes or principles. The error in this comes in forgetting the qualification that we "cannot condemn them, or find it in our hearts to maintain the superiority of our ways." When this is true, toleration is indeed the right upshot. But it is not always true.

If I go to other countries, I find other funerary practices. This might shock me, initially. But I learn that the essential human practices of dealing with death and grief and survival share their core function beneath their surface diversities. After this I cannot find it in my heart to maintain the superiority of our ways: it becomes a matter of choice whether we cremate, or bury, or leave for the vultures, and a good thing too. But if I go to Afghanistan, the situation is different. I can, and should, maintain exactly what I started with. If I become infected by the Taliban attitude to women, that is unfortunate and represents a deterioration in my own moral fibre. If the Taliban are seductive enough in other ways, I may have to be on my guard against this.

Clearly, then, although there is no moral problem of relativism there remain moral problems connected with multiculturalism. There are particular problems of when to tolerate and when to oppose, and the answers to these may not all be

easy, or all given in advance. We have to ask whether we are faced just with an alternative, equally good, solution to some problem of living that can cheerfully be acknowledged in the spirit of an alternative convention, or whether, on the other hand, we are faced with something to which we are rightly averse. Slavery, the oppressions of caste systems, the systematic degradation of women, child labor, gross inequalities, repression of opinion, and many other facets of many societies are not alternative equally good solutions to problems of living. They are things that must be opposed.

So the expressivist or quasi-realist approach gives a complete defense against relativism, acknowledging only particular problems that have to be solved, when they come up, like all moral problems. We have to approach them deploying the beliefs and attitudes that we hold, and bringing them to bear as best we can. If I am worried about whether, say, to tolerate lesbian parenthood or the use of cannabis as alternative lifestyles, or whether to oppose them as impermissible aberrations, this is what I have to do. They may get put in the same category as the divergent funerary rites, or they may go with the Taliban. But each issue has to be fought on its merits. There is no problem of relativism, but only individual problems of living.

Approaches That Have Problems with Relativism

Why do I think it an advantage of quasi-realism that it solves this issue so cleanly? Cannot other approaches say the same things? The difficulty is this. Many philosophical attempts to understand the nature of ethics look for more "robust" or "substantive" conceptions of ethical truth. Frequently they try to model ethical truth on other areas: science and a teleology for human beings in the case of some kinds of naturalism; mathematics, in the case of a priori and constructivist approaches; secondary qualities in the case of some contemporary "moral sense" theories. Such theories typically arise in response to a felt need for some substantive conception of moral truth. Now where there is a need, there is also a danger. The danger is that what you get actually fails to fill the need in the right way. It is here that relativism becomes a problem. I shall illustrate the threat with three examples.

The first example is the kind of theory associated with John McDowell that uses the rule-following considerations to defend a close analogy between ethics and secondary-quality perception. This gets into difficulties over relativism, because both the rule-following considerations *and* the analogy with secondary-quality perception encourage relativism. Roughly, in each case we can envisage a situation in which there are different "whirls of organism." There are organisms that whirl the Taliban way, and see women as inferior beings whose highest purpose is passively to serve the pleasures of men, and organisms that whirl the Western or enlightenment way. There are organisms that whirl the way dogs do when it comes

to smells, and organisms that whirl our way. If truth was found in the "responses" or the "practice" or the "shared consensus" of organisms, then it is very hard to see why these individual communities of shared response are not generating their own truths. This is how we do think of it in the case of secondary qualities. The dog inhabits, literally, a different world of smells from the human being. And there is no saying that just one of us is "right." So relativism becomes a real threat, because the theory looks as if it has to allow for a plurality of truths. There are defensive moves possible (one can always remark that one element of our whirl of organism is to set ourselves against those who whirl another way) but the danger is very real.

For a second theory that courts danger, consider the very different view of Christine Korsgaard (1996) – a kind of constructivism. Korsgaard fears the contingent, changeable world of desire and attitude enough to want an entirely different source of normativity, which she finds in such notions as self-legislation, and the nature of practical identity. Now although Korsgaard herself believes that there is a kind of Kantian straightjacket on the shape our self-legislations must take, the problem of relativism is more impressive than this solution to it. For on the face of it, you can have different groups of self-legislating persons whose identities are happily bound up in various constraints they set themselves under, but who unfortunately find these constraints in entirely different places. The Taliban way of life gives men a practical identity. It includes thinking it is their duty to privilege men over women. Similarly some people would find their identities threatened if they broke certain dietary prohibitions, while others do not care at all. If the "construction" or self-legislation generated moral truth, once more we seem to have a plurality of truths, and relativism strikes. And the same result awaits the less rationalistic, more political, version of constructivism that seemed visible in some of Rawls's later work. That is, whereas in Rawls's original work we might have thought that the framework of a decent society emerged as the necessary upshot of rational choice, in the later work we have a more relativized conception of the kinds of structure that we can justify to each other, given conditions of political dialogue that are themselves contingent and local. The shift in emphasis seems to me wholly admirable, but it leaves the door open to relativistic fears that the more ambitious Kantian coloring once closed.[4]

Finally, the same kind of result looks to threaten any neo-Aristotelian theory that seeks to tie moral choice to some conception of "flourishing." The problem is again that there is a plurality of ways in which flourishing can be had. Some flourish one way, some another. It is not just that there is, as Bernard Williams memorably put it, no particular tie between behaving admirably and flourishing "by the ecological standard of the bright eye and the glossy coat." It also remains disappointingly true that in such a flexible animal as the human being, entirely different conceptions of flourishing will support entirely different conceptions of what to do and what to admire. We have only to think of the Taliban conception of what it is in a woman to flourish (a conception, incidentally, that might be shared by the oppressed, after generations of habituation). Once more if moral

truth is found "in" such ideas, then given the plurality of ideas, we have the relativistic plurality of truths.

These views of where moral truth is found all meet difficulties of this kind. Of course, this is not to deny that conceptions of what counts as flourishing should *inform* our attitudes. But when the conceptions are contestable, we are under no pressure at all to think that they generate different and contesting systems of ethical truth. They only generate the need to choose, and to defend our choice with the best story we can find.

Objectivity and Relativism

Objectivity is very important to many philosophers of ethics. Can the view I have sketched defend a sufficiently robust concept of objectivity to satisfy us? Here too I shall urge that there is no problem.

Objectivity, I say, is desirable. It is a virtue. But what does it mean? We can think of this by considering the flaws and failures that denote its absence. First, there are flaws of *bias*. Two considerations are equal, but the biased judge weighs one more heavily than the other. Two people have equal claims, but one is preferred to the other. The most obvious cases are ones in which only a certain range of considerations *ought* to affect an issue, but others are surreptitiously introduced. Thus, I hold that some restricted range of considerations ought to influence a hiring decision. If my colleague introduces another – say, trying to reject a candidate because of age or gender or haircut – then he is not being objective. He is indulging his bias. Famously, we are not good at knowing when this is true of ourselves, and the mechanisms of self-deception are familiar enough. The colleague may not actually advance age or gender or haircut as a reason. But if, sufficiently often or sufficiently predictably, the reasons he does advance all turn out to discriminate against those of the wrong age or gender, or unusual haircuts, we know what is going on. And we may know it before he does.

Worse than bias, epistemologically at any rate, is blindness. For as well as objective decisions, we talk of objective views of things. A person may fail to give an objective appreciation of the situation because he or she fails to see the situation at all. A person who does not appreciate what is going on cannot give an objective view of a situation. To be objective, our view has to have taken enough into consideration. It has to be sensitive to what matters.

In the case of both decisions and views this formula seems to work. To be objective is to be sensitive to the right aspects of the situation, and in the right way. This means that a sense of right judgment is built into the concept at the center, and, of course, we can expect disagreements. My colleague may believe that he is *right* to discriminate as he does. That is, he has some story defending his discriminations, and he may try to get us to listen to it. Then once more we have a practical problem. We may find ourselves listening to him, or we may know

in advance that it is no use doing so. It is not given that we never budge from our antecedent opinions, but it is not given that we simply roll over and agree with his either.

Can the expressivist or quasi-realist select an appropriate range of considerations as the *right* range? Can he or she privilege the decision making of the unbiased judge, or the point of view of the informed, large-minded spectator who sees the situation in its entirety? Certainly. I am against my colleague who lets his hiring decision be influenced by the age, gender or haircut of the applicant, and I express this by saying that he is sensitive to the wrong considerations. He is biased, not objective. I am impatient with people who get their opinions from the gutter press, because that press does not give a full enough selection of facts to justify their attitudes. I privilege the better informed.

Let us return to the Taliban. Can the expressivist or quasi-realist say that there is anything *objectively* wrong about their views? They are certainly sensitive to the wrong considerations – that is, given by the fact that they let educational policies get decided by gender. I should also say that they are blind to the nature of women and the possibilities open to them. They are insensitive to most of the important aspects of women's lives. So certainly, it is a plain fact they show deficiencies of objectivity. They are objectively wrong.

What else might the worry about objectivity be? It might be a way of reintroducing the demand for proof that we have already met. Someone might, I suppose, seek to use the notion so that someone is objectively wrong if and only if there is a guarantee in advance that there is a cognitive procedure for changing his or her opinion to coincide with ours. But this would be an unfortunate usage: witness my illustration that there is no such procedure even in the straightforward empirical cases where people are happiest talking of objective error.

One thought people may have in mind, when they hanker after objectivity, is this. If we are wrong about literal objects, such as what lies in our path, we expect to be tripped up. Objects make themselves felt. The captain of the *Titanic* was wrong about an object, and as a result lost his ship. We might hope to show that if the Taliban are "objectively" wrong, then there is disaster lying in wait for them: an equivalent to being tripped up. This is one side of the thought we have already met, that associates ethical truth with flourishing, and issues in Aristotelianism. The idea will be that the objectivity of the Taliban mistake will be manifested in the relative impoverishment or loss of flourishing lives. Just as people think that the economic errors of communism were finally exhibited in the collapse of eastern European communism, so they might hold that errors of morality will be illustrated in the collapse of lives built on those errors.

If that is the idea, the expressivist tradition certainly does not have to oppose it, although I myself would not put much store by it. It *may* be true that there are such tight constraints on flourishing that any deviation from a good ethic destroys it. Or rather, we should say, it *could* have been true. For it is not really a sensible thought to keep open, after all we know about human beings. It is much better to think of the alignment between flourishing and behaving well as highly

contingent, and not only contingent, but *politically* contingent. That is, it is up to a society to create an alignment between behaving well and flourishing. As Hume (1888) argues in the third book of the *Treatise*, commercial societies implement attitudes and procedures so that people who renege on their word get tripped up. If they do not get tripped up, because, for instance, transactions are not repeated, then the alignment breaks down, and we get individual cases in which the cheat flourishes.

It would be up to the world community to ensure that a society that refuses to educate its women is tripped up – that is, somehow penalized in ways that eventually result in reform. When it comes to things like caste systems or oppression of minorities, I see no reason at all for thinking that *nature* puts into place mechanisms for tripping up those who hold the wrong attitudes. It is "set us as a task," in Kantian terms, to put into place the sticks and carrots that will hopefully introduce improvement.

Aristotelian and Kantian traditions share an overreaching ideal. The Aristotelian hopes to show that other ways of life impede human flourishing. The Kantian hopes to show that they have transgressed against some rational constraint on practical reasoning, and that this is the bottom line: this is what is wrong with them. Whereas the real truth is that what is wrong with them is neither of these things. So far as I can predict the Taliban might well flourish, again by the ecological standard of the bright eye and the bushy coat, if the rest of the world does nothing about it. They may flourish by any standards they themselves would recognize (and, as already mentioned, that could include female members of the society). And it is pie in the sky to believe that there is some theorem of practical reasoning against which they trespass, or that this is what is wrong with them. What is truly wrong with the Taliban is more straightforward than either of these things. What is wrong with them is that they oppress their women, impoverish their lives, and keep them in a state of ignorant superstition. Why should we feel any urge to say *more* than that? Isn't it bad enough?

The idea behind saying "They are objectively wrong" on this suggestion was that they are wrong in the kind of way that those who are wrong about objects are: they get tripped up. The world itself contains the mechanisms for correcting their ways, and preferably in ways that they cannot ignore. I have suggested that this is too optimistic. And again, we might reflect that it is pretty optimistic even in many empirical cases. Certainly if I am wrong about whether there is a cliff or iceberg in front of me, the world will trip me up. But if I am wrong about less immediate matters, I can go happily with my errors to the grave. If you disagree with me about some celebrated contemporary trial, or some historical fact, one of us is objectively wrong, but neither of us is particularly likely to suffer because of it. The connection between objectivity and flourishing is not so very close in everyday empirical cases, so it is unwise to ask for it to be closer in the delicate cases of ethics.

To sum up: if someone calls our opinion about an ethical issue "subjective" we can hear the charge in a number of ways that matter. She may be imputing hidden

bias. She may be imputing lack of knowledge, or lack of ability to deploy the right knowledge. She may think that the issue is genuinely one on which it is possible to be in two minds, and that any more definite attitude is only one option, or one unattractive option. All these are charges we may have to listen to, and on occasion they may be justified. But there is no single general charge, always waiting to be made. My judgment, like that of others, will doubtless show particular flaws of this kind in particular situations, and it is then up to the critic to press the particular charge he or she has in mind. And that too is just part of the human situation.

Authority: The Last Word

In his book *The Last Word* Thomas Nagel set up and considered a "relativist" challenge to the authority of norms, not only in ethics but potentially in other areas too, such as logic, mathematics, or science. His hate figure was the postmodernist or relativist who hears the right opinion, and even echoes it, but adds always the qualification or rider: that is just us. Nagel opposed to the bitter end this last word. He even invoked a kind of Platonic image of harmony between our ways and the normative orders that govern the universe, urging that if that is to be the alternative, it is preferable to the relativism, whatever metaphysical anxieties it provokes. Nagel's anxiety, and the cost he is prepared to pay for a remedy for it, has undoubtedly been widely shared. Yet the angst was unfounded, and the cost does not have to be paid. This is fortunate, for I believe it cannot be paid: there is no proof awaiting us that our normative attitudes harmonize with the normative order governing the universe, for there is no such order, and even if there were it would only filter through to us through the medium of people's consciences, and those diverge in their delivery.

Nagel was, however, right that there is something flattening or dispiriting about the relativistic last word. And it was surely intended by many postmodernists as a debunking signal: a flag showing that they had *seen through* the authority of whatever norms are in question. They may go on to *voice* acceptance of the norms, but, to Nagel's ear, the protestation will sound hollow. They cannot give the norms the *authority* they really deserve. And in Nagel's view, only the Platonist can do that.

My comment on this dialogue is once more that the mistake is one of taking there to be One Big Question, when in fact there are only a lot of little questions. Nagel thinks the relativist challenge points us toward a large metaphysical hole: a gap in our ontology or view of the world that we must desperately try to overcome. I believe, instead, that it amounts at best to an attempt to undermine our *first-order* confidences, and that such attempts are to be met piecemeal, depending upon the case in hand.

Here is an illustration from work by the late Jean Hampton. Consider the scientific norm of insisting upon double blind tests for the efficacy of new drugs (so that nobody involved knows which is a real drug and which is a placebo). Suppose some halfhearted experimentalist who says "We like to do double blind tests" and then adds "but that is just us." Surely the right response is to ask what the little word "just" is doing. It insinuates the double blind methodology is *optional*, so that there would be nothing wrong about doing the tests without it. But that, we suppose, is a shocking mistake. Tests conducted without it have a high probability of saying that a drug is efficacious when it is not, or vice versa, and this is what we wanted to avoid. If our experimentalist denies this, he had better have a story, and most scientists would bet in advance that he will not have one.

Another good way of putting it is this. Let us give the dissident that it is just us. But let us add some words of self-awareness. It is just us, able as we are to conduct reliable tests. Or, it is just us, forearming ourselves against misleading results. There is nothing to be *ashamed* of in these words of self-description. On the contrary, there is something to be *proud* of, because learning to avoid the experimental pitfalls presumably took some doing. So we can agree with Nagel that better last words than "that is just us" should be found. But we stay within our first-order normative space as we find them. This makes it plain that the issue with the relativist is fundamentally one of confidence. But if our last words are happy enough, our confidence is well founded, and no challenge to it exists.

My response is the same if an ethical attitude is challenged. I am in favor of education for women. Suppose now not a Taliban member, but some weary postmodernist, saying, "I am in favor of it too. But that is just us." Again the "just" insinuates that this is somehow an optional attitude; that there is *nothing wrong* with people such as the Taliban who happen to whirl the other way. But I am unsympathetic to this degree of toleration – the kind of open mindedness that comes when all one's brains have fallen out, as it is sometimes put. Putting it positively I can add words of self-description: it is just us, free from the politics of arbitrary discrimination, free from culturally embedded misogyny, maintaining ideals of equality and freedoms of self-development for all. These words ring well in my ear: it is just us, and we are doing well. Putting it negatively, I can expound as I have done on the way the Taliban attitude is not optional. There is indeed something wrong with it and them, and I can go on to detail some of what it is.

Of course, in all this I am speaking in my own voice. That is inevitable. But so long as my own voice is not one to be embarrassed by, it is also of no interest.

Notice that this is not an open invitation to be smug. Suppose I announce my hostility to legalized cannabis. Someone comes along and, perhaps pointing to countries where cannabis is legal and things seem to go on fine, says, "We may hold it right to criminalize cannabis, but that is just us." I try for words of self-description: just us, aware of the dangers; just us, able to discriminate harmless

drugs from harmful ones. And here, suddenly, the words may not ring all that happily. Either the descriptions are unjustified or false, or they are not things to be proud of. They may even be things about which to be embarrassed: "Just us, ignorant of the real consequences" or "Just us, believing what we are told in the gutter press." When this happens something has to give. We either have to find better words, or change our habit of being embarrassed by ignorance or gullibility, or come to soften our opposition to cannabis. Or, perhaps, "carelessness and inattention" can provide a remedy, for one human response to finding that the last words of self-description are unflattering is, unfortunately, just to kick over the table, refusing to listen. But that too is an embarrassing last description of our state.

We can conclude by noticing one final issue that perhaps troubles people who think about these things. I suspect it lies behind the kind of moral imperative people feel to transcend quasi-realism.

As we have seen, people hanker after algorithms, or procedures guaranteed in advance to prove that their opponents are wrong or, preferably, to prove to them that they are wrong. They want this security. But suppose, as I am afraid is true, they cannot always find such procedures. Well, in fact, attitudes and politics are going to continue. In practice, however much professors beat their breasts about the need for an algorithm in the classroom, out of the study they will guide their children, sit as magistrates, campaign for one party or another, lament the Taliban, just as if they had one. But one strand from Kant will lead the professors to be embarrassed about some of this. For the attitudes and the politics will eventually have a coercive edge. There will come a hard case, where all they can do is withdraw goodwill, or shun the malefactor. And people think that this trespasses against the *respect* due to persons, or against the *dignity* of persons – and this worries them. People thinking like this fear that they can only be justified in coercive measures if, at least in principle, the object of those measures ought to be assenting to them, and this in turn would only be so if somewhere there is the missing algorithm. It is as though, when I lock you up, if what I do is just, then you must be tacitly in possession of thoughts that would lead to you assenting to your fate. You could reason to my conclusion, even if in fact you choose not to. Otherwise, so the line goes, I fail to respect you as a person, or infringe your dignity and even your rights.

But sad experience denies the dream that at the end of every just piece of sentencing there is a prisoner who tacitly accepts the justice of the sentence. We may desire that the malefactor repent, recognize the wrong he or she has done, reform and rejoin the rest or us. Adam Smith (1759) thought that this was the true end or goal of resentment. But human beings do dirty things, and their cultures may have taught them to do them. If a Taliban member living in a Western jurisdiction is indicted for the not-uncommon crime of burning his wife to death for some petty shortcoming or misdemeanor, he may feel unjustly treated. His whole culture may back him in this resentment. He may feel it atrocious that his crime attracts the same penalty as if he had killed a male relative.

And he may have no principles within him that could even ideally be deployed to change his feelings. But that does not give me, the Westerner, the least reason either for changing the judgment, or for feeling guilty about enforcing a serious penalty. Nor do I trespass against dignity and due respect. The Taliban member forfeited that. He deserves no respect for his attitude to women or for his action. Nor does his self-justification express any kind of dignity. He brought it all upon himself, by himself refusing to treat half of humankind with due dignity and respect.

A weaker and more plausible requirement is that coercion should only be exercised against someone if it would be *reasonable* to expect people to agree that this is a case where coercion is necessary. The agreement sought after would not necessarily be that of the criminal, for as we have already described, the criminal may be blind to what shows the error of his or her ways. But so long as reasonable people see the error of the criminal's ways, coercion may be justified. This seems to be right, but it really amounts to little more than an equation between behavior being justified and it being seen by reasonable people that it is justified. "Reasonable" is here functioning in what we might call a Hume-friendly sense. It does not mark the delivery of pure practical reason, uncontaminated by our contingent concerns, and sympathies and attitudes. It means an appeal to people who are objective, in the sense described earlier, and who, having thought the matter through, sifted and refined their attitudes, decide that the criminal should indeed be subject to coercion. Perhaps it does trespass against the dignity or respect due to a person if we coerce him or her although our policy does not meet this condition. But that means only that we trespass against these things when our policy is unjustified. I am quite prepared to believe that this is true, but it is not as substantive a truth as we might have hoped to find.

It would, of course, be nice if all criminals were penitents, and if the dignity of people and the respect due to them were always so great that there was no need for coercive measures in the first place. And no doubt these things are true in fairyland. In the real world, we have to do the best we can, and here, as elsewhere, the task of the philosopher is partly to enable us to do this without the confusions and guilt that so often bedevil practical reasonings.

Notes

1 Most famously, at Wittgenstein (1953).
2 Some writers make a distinction whereby the moral is a subset of the ethical. Ethics is concerned with the whole field of evaluation, choice, and action, whereas morality is concerned more with obligations and duties. In this essay the distinction is not important, and I shall talk indifferently of the moral and the ethical.
3 For more on Ramsey's ladder, see Blackburn (1998).
4 The major works contrasted here are of course Rawls (1971) and Rawls (1993).

References

Blackburn, S. (1998) *Ruling Passions*, Oxford: Oxford University Press, pp. 78, 294–7.

Hume, David (1888) *Treatise of Human Nature*, ed. L.A. Selby-Bigge, Oxford: Oxford University Press, III, ii, 5, p. 516ff.

Korsgaard, Christine (1996) *The Sources of Normativity*, Cambridge: Cambridge University Press.

Mackie, John (1977) *Ethics: Inventing Right and Wrong*, Harmondsworth, UK: Penguin Books.

Rawls, J. (1971) *A Theory of Justice*, Cambridge, MA: Harvard University Press.

Rawls, J. (1993) *Political Liberalism*, New York: Columbia University Press.

Rorty, Richard (1989) *Contingency, Irony and Solidarity*, Cambridge: Cambridge University Press, pp. xv, 173.

Smith, Adam (1759) *The Theory of Moral Sentiments*, II, iii, I, I, 5.

Strawson, Peter (1968) "Freedom and Resentment," in *Studies in the Philosophy of Thought and Action*, Oxford: Oxford University Press, p. 79.

Wiggins, David (1990/1991) "Moral Cognitivism, Moral Relativism, and Motivating Beliefs," *Proceedings of the Aristotelian Society* 91: 61–86.

Williams, Bernard (1985) *Ethics and the Limits of Philosophy*, London: Fontana.

Wittgenstein, Ludwig (1953) *Philosophical Investigations*, Oxford: Blackwell, §136.

Chapter 3

Moral Agreement[1]

Derek Parfit

The Argument from Disagreement

We cannot rationally believe that there are moral truths, it is often argued, given the facts of deep and widespread moral disagreement, and the cultural origin of many moral beliefs.

To introduce this argument, I shall sum up some claims that I defend elsewhere (Parfit 2011: Part 6).

(A) There are some irreducibly normative reason-involving truths, some of which are moral truths.
(B) Since these truths are not about natural properties, our knowledge of these truths cannot be based on perception, or on evidence provided by empirical facts.
(C) Positive substantive normative truths cannot be analytic in the sense that their truth follows from their meaning.

Therefore

(D) Our normative beliefs cannot be justified unless we are able to recognize in some other way that these beliefs are true.

We do, I believe, have this ability. We have reasons to have certain normative beliefs, and we can respond to these reasons. Normative beliefs can also be self-evident and intrinsically credible. One example is

The Blackwell Guide to Ethical Theory, Second Edition. Edited by Hugh LaFollette and Ingmar Persson.

(E) Torturing children merely for fun is wrong.

There are similar nonnormative beliefs, such as

(F) No statement can be both wholly true and wholly false.

Since our normative beliefs are neither caused by what we believe, nor based on empirical evidence, we need another word to refer to our way of forming these beliefs. On the view that I have called

> *Intuitionism*: We have *intuitive* abilities to respond to reasons and to recognize some normative truths.

Though it is intuitively clear that certain acts are wrong, most of our moral beliefs cannot depend only on such separate intuitions. We must also assess the strength of various conflicting reasons, and the plausibility of various principles and arguments, trying to reach what Rawls calls *reflective equilibrium*. This kind of intuitively based reflective thinking is not only, as Scanlon writes, "the best way of making up one's mind about moral matters . . . it is the only defensible method" (Scanlon 2003: 149).

We have similar abilities to recognize truths about what is rational, and about what we have reasons to believe, want, and do.

Many recent writers reject such claims. Schiffer (2003: 252), for example, doubts that moral intuitions are worth discussing, and Field (2000: 119–20 n. 6) and Boghossian (2000: 231) call the idea of rational intuition "obscurantist" and "a mystery." But these criticisms are aimed at the view that intuition is a special quasi-perceptual faculty. That is not the view that I am defending here. When I use the word "intuitive," I mean what Boghossian means when he describes some of his claims as "intuitively plausible" and "intuitively quite clear" (Bogossian 2008: 223, 250).

Intuitionism can also be challenged with claims about disagreement. When Boghossian denies that beliefs can be intrinsically credible, or self-evident, he points out that

(G) different people might find conflicting beliefs self-evident. (Boghossian 2000: 239)

If we claim that we have some ability, however, it is no objection to this claim that we might have lacked this ability. Different people might often have conflicting visual experiences, which were like dreams and hallucinations, and were not a source of knowledge. But that is not in fact true. Different people's visual experiences seldom conflict, and believing what we seem to see is a fairly reliable way of reaching the truth. It may be similarly true that, after careful reflection, different people would seldom find conflicting beliefs self-evident. Believing what seems

self-evident, after such reflection, may be another fairly reliable way of reaching the truth.

When Schiffer argues that there are no moral truths, he claims that

(H) even in ideal conditions, when everyone knows the relevant facts and is reasoning equally well, we and others could rationally disagree about any moral question.

For example, Schiffer claims that, though we could rationally believe that

(E) torturing children merely for fun is wrong,

it would be equally rational to reject this belief.[2] This argument assumes that we cannot have decisive reasons to have our moral beliefs. If we had such reasons to believe (E), it would not be equally rational either to have or to reject this belief. What Schiffer calls his *error theory* might be true, since we might never have decisive reasons to have any moral belief. But Schiffer cannot support this theory by claiming that we and others could rationally disagree about any moral question, since this claim assumes that we have no such reasons. Nor could we refute Schiffer's theory merely by claiming that we and others could *not* rationally disagree. When we are trying to decide whether we have decisive reasons to have certain beliefs, we cannot usefully appeal to claims about whether, when considering these beliefs, we and others could rationally disagree.

There is another way to challenge Intuitionism. Rather than claiming that we and others *might* disagree about normative questions, or *could rationally* disagree, Anti-Intuitionists might claim that

(I) even in ideal conditions, we and others *would in fact* disagree.[3]

These people might then argue:

> Since there would always be such normative disagreements, we cannot justifiably or rationally believe that our normative beliefs are true, nor can we rationally believe that any normative beliefs might be true.

We can call this *the Argument from Disagreement*. If (I) were true, this argument would have great force. If we had strong reasons to believe that even in ideal conditions we and others would have deeply conflicting normative beliefs, it would be hard to defend the view that we have the intuitive ability to recognize some normative truths. We would have to believe that when we disagree with others it is only we who can recognize such truths. But if many other people, even in ideal conditions, could not recognize such truths, we could not rationally believe that we have this ability. How could *we* be so special? And if none of us could recognize such normative truths, we could not rationally believe that there *are* any such truths.

To answer this argument, Intuitionists must defend the claim that in ideal conditions we and others would not have such deeply conflicting beliefs. According to what we can call this

> *Convergence Claim*, or *CC*: If everyone knew all of the relevant nonnormative facts, used the same normative concepts, understood and carefully reflected on the relevant arguments, and was not affected by any distorting influence, we and others would have similar normative beliefs.

Unlike the claims that different people might disagree, or could rationally disagree, CC is an *empirical* claim. Though it is a normative question what would count as ideal conditions, it is a psychological question whether in these conditions people would have similar normative beliefs.

When Intuitionists claim that we have intuitive abilities to respond to reasons and to recognize some normative truths, they should admit that we are fallible. Even in ideal conditions some people would make mistakes, and there would be some disagreements. There may be some normative questions about which, given our present abilities, we would all make mistakes. To answer the Argument from Disagreement, it would be enough to defend the prediction that in ideal conditions we would *nearly* all have *sufficiently similar* normative beliefs. Even mathematicians sometimes disagree, but they can recognize mathematical truths. We may also make mistakes about whether and when the ideal conditions have been met. There may be relevant facts or arguments, or distorting influences, of which we are not yet aware. Our normative thinking is still in its childhood.

For CC to be a significant claim, our concept of a distorting influence must be purely procedural. When someone's normative beliefs have been influenced in some way, we should not claim this influence to be distorting merely because it leads this person to have some normative belief that we reject. That would make it trivial to claim that if no one was affected by any distorting influence we and others would not disagree. We must have other reasons to believe that an influence of some kind is likely to distort our own and other people's normative beliefs. One such distorting influence would be our knowledge that if other people accepted and acted on some normative belief this would give special benefits to us.

In trying to decide whether CC is true, we must consider various historical and psychological questions. We must ask how much and how deeply people have disagreed, and how such disagreements can be best explained. We cannot hope to reach more than very partial answers to these questions. Given these answers, we must then try to predict whether in ideal conditions these disagreements would be sufficiently resolved.

In asking whether the Convergence Claim is true, I have just said we cannot appeal to our own normative beliefs. We should also set aside our metaethical or metanormative beliefs. Unlike most of us, for example, Schiffer denies that

(E) torturing children merely for fun is wrong.

But this disagreement does not count against the Convergence Claim. Schiffer calls such acts *abhorrent*, and he rejects (E) only because he believes that there are no moral truths. Schiffer would agree that if there were any moral truths (E) would be one such truth. Schiffer also calls it puzzling that rational adults use the concept *morally wrong,* since he believes that moral beliefs are best regarded as one kind of desire. Many people make such claims. When we ask whether in ideal conditions we would all have similar moral beliefs, we should use the phrase "moral belief" in a metaethically neutral sense, which allows that such beliefs might merely be, or be expressions of, such moral desires or sentiments. Like such other Sentimentalists as Hume and Blackburn, Schiffer has what are close enough to moral intuitions.

We cannot assume that everyone has such moral beliefs, sentiments, or intuitions. That seems not to be true of those who are now called *psychopaths* or *sociopaths.* On one estimate, the proportion of such people is 1% of women and 3% of men. Since these people seem not to have moral beliefs or intuitions, we cannot claim that in ideal conditions their moral beliefs would be similar to ours. But this fact does not threaten the claim that we have the intuitive ability to recognize some moral truths. That claim does not apply to people who have no moral beliefs or intuitions. Most of us can see, though some of us are blind.

Intuitionism *would* be challenged if it were true that even in ideal conditions there would be many deep disagreements between people who clearly *do* have moral beliefs, sentiments, or intuitions. There would be no such disagreement about the wrongness of torturing children merely for fun. But there are many other, more controversial moral questions. Intuitionists need not claim that in ideal conditions these disagreements would all be completely resolved. But they must defend the claim that there would not be deep and widespread moral disagreements.

The Convergence Claim

When we discuss normativity, it is a mistake to consider only morality. So we can first ask whether, as intuitionists claim, we can recognize some epistemic normative truths. I believe that

(J) when some fact implies that some belief must be true, this fact gives us a decisive reason to have this belief.

Though Schiffer denies that there are any moral truths, he accepts (J). Schiffer calls (J) "about as analytic as anything can be" (Schiffer 2003: 263). The truths that are most analytic are those that are true by definition. Schiffer's examples are

(K) Every widow was once married,

and

(L) We ought not to do what is wrong. (Schiffer 2003: 249)

If (J) were like (K) and (L), by being true by definition, we could not appeal to (J) in defending the Convergence Claim. (J) would not then be a substantive normative truth, but what Schiffer calls a "trivial truism."[4] Nor could our belief in (J) help to show that we have the intuitive ability to recognize some normative truths. To recognize that some claim is true by definition, we do not need any normative intuition.

On some uses of the phrase "a reason," (J) may be true by definition. But I use (J) as a normative claim, which could be restated as

(M) when some fact implies that some belief must be true, this fact gives us a decisive epistemic reason, by counting decisively in favor of our having this belief.

This truth, I believe, is very different from trivial truths like (K) and (L). To explain how those other claims are true, it is enough to say that the word "widow" means "a woman who was married to someone who has died," and that in saying that we ought not to do something we mean that this act is wrong. We cannot similarly claim that when we say that

(N) some fact implies that some belief must be true

we mean that

(O) this fact counts decisively in favor of our having this belief.

(N) and (O) have quite different meanings. (N) is not a normative claim. (M) states the substantive normative belief that (N)'s truth would make (O) true.

When Schiffer discusses truths like

(K) Every widow was once married,

he calls these truths "*conceptual* or *concept-based* in the sense that no one could fully understand these claims without believing that they are true" (Schiffer 2003: 249).

Schiffer might claim that (M) is also a conceptual truth. This use of the word "conceptual" may be misleading. If we could not fully understand some claim without believing that this claim is true, the explanation may not be that this claim's truth is based on the concepts with which this claim is stated. We might be unable to disbelieve such a claim because this claim is so obviously true.

(M) is not, however, a conceptual truth in Schiffer's sense. I believe that

(M) when some fact implies that some belief must be true, this fact gives us a decisive epistemic reason, by counting decisively in favor of our having this belief.

But since (M) is an irreducibly normative claim, (M) states an irreducibly normative, nonnatural truth. Some Metaphysical Naturalists understand (M) but reject this claim, because they believe that there cannot be any such truths.[5]

In asking whether the Convergence Claim is true, we should set aside such metaethical disagreements. Though I believe that (M) states an irreducibly normative truth, we should ask whether everyone would accept (M) understood in a vaguer, metaethically neutral sense. The answer, I believe, is *Yes*. When so understood, (M) is a substantive normative claim that, in ideal conditions, everyone would accept.

It might be objected: "(M) cannot state a *normative* truth. Norms must be able to be breached or contravened. It would be impossible to know that some belief must be true without also having this belief." This objection is, I believe, mistaken. People sometimes know that some belief must be true without really believing this truth because they continue to think and act as if this belief were false. It is a normative claim that what these people know gives them a decisive epistemic reason to have this belief. My claims about (M) could, however, be applied to other normative epistemic truths. One example is:

(P) If we know that, given what we know, there is a chance of 99 in 100 that some belief is true, this fact gives us a strong epistemic reason, by counting strongly in favor of our having this belief.

(P) is a substantive normative claim that in ideal conditions nearly everyone would accept.

Consider next:

(Q) The nature of agony gives us a reason to want to avoid future agony.

This claim is not, I believe, a conceptual truth. It does not follow from the meaning of the word "agony" and the phrase "a reason" that we have such an object-given reason to want to avoid agony. (Q) is another example of an intuitively recognizable normative truth. I believe that, as Nagel claims, (Q) is intrinsically more plausible than any argument that we might give in (Q)'s defense. Many people either do not have the concept of a purely normative reason, or believe that there could not be any such reasons, or normative truths.[6] But if we set aside such metaethical disagreements, and another distorting influence to which I shall return, few people who understood (Q) would seriously doubt that they have such a reason to want to avoid being in agony.

We can now turn to moral disagreements. When we discuss moral beliefs, we cannot hope to show that the Convergence Claim is true. Nor, however, could skeptics show that this claim is false. We can reasonably predict or hope that in ideal conditions we would nearly all have sufficiently similar moral beliefs. Though there have been many moral disagreements, most of these disagreements do not, I believe, count strongly against this prediction. In most cases, some of the ideal conditions are not met.

First, when different people have conflicting moral beliefs, that is often because these people have conflicting nonmoral beliefs, or because they do not know all of the relevant nonmoral facts.

Some examples are disagreements about distributive justice. There have been many conflicting beliefs about people's property rights, or the inheritance of wealth, or whether some people ought to be paid much more than others, or about which areas of land, natural resources, or man-made goods ought to be privately or publicly owned. These disagreements are often ignored by moral theories. But compared with many questions about which acts are right or wrong, such as questions about when it is right to lie or break some promise, it is more important to ask which inequalities in wealth and income can be morally justified. These inequalities have much more significant effects on people's lives. Disagreements about these questions often depend on people's having conflicting beliefs about human nature, and about the likely effects of different policies or institutions. Similar remarks apply to many other moral disagreements, such as many disagreements about sexual morality, or about our obligations to our close relatives, or about which acts should be illegal and when and how people ought to be punished. When such disagreements depend in part on conflicting nonmoral beliefs, it may be true that if we all knew the relevant nonmoral facts we would come to have similar moral beliefs.

Many other moral disagreements depend on people's having conflicting religious beliefs. Such disagreements cast little doubt on the Convergence Claim. Most of us would agree, for example, that if the Universe was created by an omniscient, omnipotent, and wholly good God, we ought to obey this God's commands.

In many other cases, our moral beliefs are affected by distorting influences. That is often true when we have conflicting interests. If we ask whether people should be paid much higher salaries when their innate abilities make them more productive, our answer may depend on whether we ourselves have such abilities. If we ask how much of their income the world's rich people ought to give to those who are poor, our answer may depend on whether we are rich or poor. When our moral beliefs are affected by our knowing such facts about ourselves, we are more likely to make mistakes. These facts ought not to influence us, since they are irrelevant to the truth of these moral beliefs. There are other distorting influences. Many disagreements cannot be ended, for example, because some people become committed to their beliefs, and are unwilling to admit that they have been mistaken.

In another large class of cases, moral disagreements are superficial, since they are about different ways of applying some more fundamental principle. When Mackie defends his error theory, he appeals to the fact that people in some societies believe in monogamy, but people in others believe in polygamy. This disagreement is not disturbing. Consider next the belief that parents have special obligations to care for their children. Since this belief is almost universal, it does not support the Argument from Disagreement. But even this belief is not, for most of us, fundamental. That is shown by how we would respond if we considered those actual or imagined communities, such as some Israeli kibbutz or Plato's Republic, in which children are communally reared. We would not believe that in such communities parents were simply acting wrongly in failing to care for their own children. Most of us would believe that (1) people ought to play their part in whatever in their society is the established system of bringing up the next generation, and that (2) in the best system parents would care for their own children. Any disagreements about (2) would mostly depend on people's having conflicting nonmoral beliefs.

Sidgwick suggests another example which may lead us to doubt whether it is a fundamental truth that parents ought to give priority to their own children. He writes:

> If, however, we consider the duty of parents by itself, out of connection with this social order, it is certainly not self-evident that we owe more to our own children than to others whose happiness equally depends on our exertions. To get the question clear, let us suppose that I am thrown with my family upon a desert island, where I find an abandoned orphan. Is it evident that I am less bound to provide for this child as far as lies in my power, the means of subsistence, than I am to provide for my own children?
>
> (Sidgwick ME: 346–7)

There are other ways in which people may only seem to disagree. In some cases people use words like "ought" and "wrong" in different senses. Sidgwick, for example, claims that he ought not to prefer his own lesser good to the greater good of others. This may suggest that, on Sidgwick's view, he would be acting wrongly if he saved his own life rather than the lives of several strangers. Most of us would reject that view. But Sidgwick seems to be using "ought" in what I call its *impartial-reason-implying* sense. He seems to mean that, if he assessed his reasons from an impartial point of view, he would have more reason to prefer that more lives be saved, and more reason to prefer any other greater good. We would not reject that claim.

Some other moral disagreements are not about *which* acts are wrong, but about *why* these acts are wrong, or what *makes* them wrong. Different answers are given by different systematic theories, such as those developed by Kantians, Contractualists, and Consequentialists. Such disagreements do not directly challenge the view that we are able to recognize some moral truths. In defending this view, it is

enough to defend the claim that in ideal conditions there would be sufficient agreement about which acts are wrong. Though we also have intuitive beliefs about *why* many acts are wrong and about the plausibility of different systematic theories, we would expect there to be more disagreement about these other questions. As I have also argued, however, when the most plausible systematic theories are developed further, as they need to be, these theories cease to conflict (Parfit 2011: Part 3). If that is true, these theoretical wars would end.

Many other disagreements are about borderline cases. Such disagreements do not count against the view that there are some moral truths. Even when we all agree that acts of some kind are wrong, we should expect that we would sometimes disagree about which acts are of the relevant kind. We may agree, for example, that it is wrong to kill innocent human beings, but disagree about the status of a human embryo or fetus. There are two main ways in which we can use the phrase "a human being." In one use, a fertilized ovum counts as a living member of the species *Homo sapiens*, and is therefore a human being. This is like the claim that when the first green shoot emerges from an acorn, this acorn is already an oak tree. We may instead use different concepts of a tree and a human being, claiming that such a sprouting acorn is not yet an oak tree, and that a fertilized ovum or embryo is not yet a human being.[7] When people's concepts differ in this way, that may lead them to disagree about the wrongness of abortion. It is a difficult question whether and how this disagreement could be resolved. But since this disagreement is about borderline cases, it does not cast doubt on the view that it is wrong to kill innocent human beings. There are similar disagreements about which acts count as killing someone, or merely as a failure to save someone's life.

These cases illustrate another kind of disagreement. When we ask whether acts of some kind are wrong, many people assume that the answer must be all-or-nothing. In many cases, however, the morally relevant facts are matters of degree. If an embryo or fetus turns slowly into a human being, the moral objection to an abortion may similarly grow in strength. Nor should we give equal weight to the saving of each person's life. Compared with giving someone fifty more years of life, it is very different to give someone else only a single extra month, or one extra week, or day. Return next to the question of what we rich people ought to give to those who are very poor. If we assume that wrongness is all-or-nothing, we shall be most unlikely to agree on how much we ought to give. And it is hard to believe that there could be a definite answer here, so that what is wrong might be giving less than a tenth of our income, or less than a fifth, or less than half. For most of us, the truth is rather that we shall be acting less wrongly the more we give. When people have conflicting moral beliefs because they mistakenly assume that wrongness cannot be a matter of degree, these disagreements do not count against the Convergence Claim. If these people gave up this assumption, that would end such disagreements.

Many people also fail to see that in many cases normative truths are imprecise. One example is the question of how it would be best for someone's life to go. When we are making decisions that will greatly affect the rest of our lives, such as

choosing between two possible careers, or deciding whether to have children, the truth is often that neither of these possible futures would be better for us, or would make our lives more worth living. We should not assume that, when neither of two possible lives would be better, these lives must be precisely equally good. Two very different lives could not, I believe, have such precisely related values. These lives would be only *imprecisely equally good*, and this imprecision would often be great. Similar claims apply when we ask which people are worse off than others, in morally relevant senses. People in very different circumstances could not be precisely equally well off. But these questions have answers, since some lives are more worth living, and some people are better off than others. These differences are matters of degree. One life might be *somewhat* better than another, which is *much* better than a third, and one of two people might be either *somewhat* worse off, or *much* worse off. Such comparisons involve what we can call *imprecise cardinal comparability*.

It is easy to think about such cases in ways that lead us astray. When some things can be better or worse than others, and by more or less, it is natural to use what we can call *the Linear Model*. The goodness of these things, we may assume, involves a *dimension*, which we can think of as if it were a line, or scale of value. Something's goodness corresponds to its position on this line. Suppose next that, of two things, X is now worse than Y, and is therefore lower down on the line that represents our scale of value. X starts to get better in some gradual way, and ends up higher on this line than Y, thereby being better than Y. If that is how we think about such cases, we cannot help believing in precision. Since X has moved up this line from being lower than Y to being higher, there must have been a time when X was at the same point as Y, thereby being precisely equally good. In most important cases, that conclusion would be false. Suppose, for example, that X and Y are Shakespeare's drafts of two new plays. Because Shakespeare knows that one of these drafts is worse, he rewrites more than a thousand lines, thereby turning the worse play into the better play. There would be no point during this rewriting when these two plays were precisely equally good.

To understand these cases, we must reject this Linear Model, which unavoidably implies precision. Nor should we think in terms of numbers, since these would also imply precision. It would not be enough to use the idea of a *range* of value by saying, for example, that rather than having a value of 90, something's value ranges from 85 to 95. Such a thing would be only slightly worse than something else whose value ranges from 86 to 96. When we think about cases that involve imprecise cardinal comparisons, we should deliberately avoid thinking in either spatial or numerical terms – except as a form of shorthand that we should remember to be seriously misleading.

A scientific analogy may be helpful here. Before Einstein's great discoveries, many people thought of time as if it were a line, with each moment having some position on this line. On this view, if neither of two events occurs before the other, these events must be simultaneous. No third possibility makes sense. Einstein discovered that, given the surprising ways in which time is related to space and to

the speed of light, we must cease to think of the different moments of time as if they all had some position on a single line. When two events occur in sufficiently distant places, if neither event occurs before the other, this does not imply that these events are simultaneous. These events are related in a third way, which is sometimes called being in each other's *elsewhere*.

This analogy is only partial, since Einsteinian space–time involves relations that are precise. But it may help to remember the fact that for many centuries it seemed to many people to be certain that time could be represented as a line. This assumption, we have learnt, was a mistake. It may now seem similarly certain that when some things can be better than others, and by more or less, such differences in value can be represented as if they involved different positions on a line, or scale of value. When such differences are imprecise, as they very often are, this assumption is also a mistake.

It is sometimes claimed that to persuade people that differences in value can be imprecise we can show these people that they already recognize this truth in making some of their decisions. Suppose that you have been offered two jobs, *A* and *B*, which would involve very different kinds of work and would involve living in very different cities. You find it hard to choose between these offers, which seem to you equally good. The salary for job B is then significantly raised, making this offer seem much better than it was before. But this improvement doesn't solve your problem, since you still find it hard to choose between job A and this better version of B. Your continuing indecision may seem to show that you earlier believed that jobs A and B were only *imprecisely* equally good. It may seem that if you had earlier believed that A and B were precisely equally good you would have decided that this improved version of B must be better than A. But this reasoning is mistaken. You may have earlier assumed that though there must be some precise truth about the relative goodness of A and B, you knew only very roughly how good these jobs would be. That would be enough to explain how when B is improved, that does not solve your problem. This improvement may be well within your assumed margin of error.

There are other ways to defend the view that there can be such imprecise differences in value. Consider first comparisons of a different kind. Suppose that someone asks whether Einstein or Bach was a greater genius, or had greater achievements. We may think this a pointless question, since we cannot possibly compare the greatness of scientists and composers, or their achievements. But this response would be a mistake. Einstein was clearly a greater genius than any untalented fifth-rate composer, and Bach was clearly a greater genius than any incompetent fifth-rate scientist. As this shows, there *are* truths about the relative greatness of scientists and composers, and their achievements. If we had earlier believed that there could not be *any* such truths, it would be implausible to move now to the opposite extreme, believing not only that there are such truths, but also that such truths must be precise. Given the very great differences between music and physics, it could not be true, I believe, that Bach and Einstein, or their achievements, were precisely equally great. Nor could it be true that either was slightly greater than

the other. Though there can be differences in the greatness of achievements of such very different kinds, these differences must be imprecise.

Since these claims about greatness are evaluative, we can next point out that there is similar imprecision in many nonevaluative and nonnormative facts. If we are comparing two very different pieces of mechanical equipment, for example, there may be no precise truth about which of these pieces of equipment is more unwieldy, or awkward to use. And there would often be no precise truths about which of two rooms is more untidy, or which of two theories is more complicated, or which of two mountains it would be harder to climb.

Similar claims apply to the goodness of outcomes. Suppose we believe that it would be in one way better if some group of people received a greater sum of benefits, and in another way better if these benefits were more equally distributed between these people. There would often be no precise truths either about which of two sums of benefits would be greater, or about which of two patterns of distribution would be less unequal. Nor could there be precise truths about the relative importance of how great the sum of benefits would be, and how equally these benefits would be distributed. In such cases the truth would often be that (1) neither of two outcomes would be better, and that (2) these outcomes would be very far from being precisely equally good. Though we can call such outcomes equally good, it is clearer to say that neither would be better.

Similar claims apply to questions about the wrongness of acts, and about what we ought to do, or have most reason to do. There are often no precise truths either about which acts would do more good, or about the relative importance of other moral considerations or reasons for acting. There are no such truths, for example, about the relative strengths of our reasons to keep some promise, or to help some stranger who is in distress.

When different people have conflicting beliefs about which of two outcomes would be better, or which of two acts would be wrong, that is often because these people mistakenly assume that such normative truths are more precise than they really are. If these people realized that many such truths are very imprecise, they would often cease to disagree. These people would come to see that neither of two outcomes would be better, or that neither of two acts would be wrong.

There is another way in which these facts about imprecision support the view that there are some normative truths. If such truths had to be precise, it would often be hard to believe that there *are* such truths. It would be hard to believe, for example, that one of two possible lives could be 23.7% more worth living, or that one of two people could be, in some morally relevant sense, 3.16 times better off. When we see that such truths would be very imprecise, it is easier to recognize that some lives are more worth living than others and that some people are better off.

We can next briefly consider another similar, but more puzzling kind of case. Some questions may be *indeterminate*, in the sense that they have no answer. That is sometimes true, for example, of the question, "Is he bald?" If some man has no hair, he is bald. If some man has a full head of hair, he is *not* bald. But we cannot

plausibly assume that in all cases between these two extremes any man must either *be*, or *not* be, bald. In many cases, though it is not true that some man is bald, it is also not true that this man is *not* bald.

Similar claims might apply to normative questions. One example is the wrongness of abortion. Suppose that

(R) it is not true that there is any moral objection to early abortion.

This may seem to imply that

(S) there is no moral objection to early abortion.

But that may not be so. When it is *not* true that some man is bald, we cannot conclude that it *is* true that this man is *not* bald. In the same way, we might be right to believe both (R) and

(T) it is not true that there is no moral objection to early abortion.

It might not be true either that there *is* a moral objection to early abortion, or that there *isn't*. There are other difficult moral questions, such as questions about what would be overpopulation, or about the morality of war, which may have no answer.

It may seem a trivial fact that when we ask whether someone is bald, this question may have no answer. But when we ask normative questions, this possibility can be more puzzling, and disturbing. We may find it hard to give up the assumption that if it is not true that some act is wrong, this act must be morally permitted. We may think that if it *isn't* true that some act is wrong, it must be true that this act *isn't* wrong. But if every act must either *be*, or *not* be, wrong, must it not be similarly true that every man must either *be*, or *not* be, bald? And that is not true.

Cases of this kind raise several difficult questions, which partly overlap with questions about imprecision.[8] It is sometimes claimed, for example, that indeterminacy is entirely linguistic or conceptual. On this view, though our words or concepts may be vague, reality could not be vague, and we could always make our concepts more precise so that we could give fuller descriptions of the facts. But this view is too simple. There are indeed many cases of this kind. There are always precise truths, for example, about how many hairs there are, at any time, on some man's head. It is the concept *bald* that is vague, and we could introduce a precise concept, which referred to these numbers of hairs. Questions that used this revised concept might all have answers. But there are many other cases to which this view does not apply. In such cases, there is no acceptable way of making some concept precise, since such revised precise concepts would lead us to draw distinctions and make claims which don't fit the facts. These concepts and claims would treat these facts as being more precise than they really are. This is Sen's

objection, for example, to all of the criteria that economists have proposed about the relative badness of different patterns of economic inequality (Sen 1973). Similar remarks apply to claims about which lives are more worth living, or about the relative strength of many conflicting reasons. If we tried to make such claims more precise, that would often make these claims false. As before, similar remarks apply to nonnormative claims. We might, for example, truly claim that one of two theories was about twice as complicated, or that one of two mountains was about twice as hard to climb. But if we said that this theory was 2.17 times as complicated, or that this mountain was 2.17 times as hard to climb, these claims could not possibly be true.

There are also some powerful arguments against most accounts of indeterminacy. We may assume, for example, that if some man is not bald, no removal of any single hair could make this man bald. But that seems to imply that even if we removed every hair from this man's head, one by one, we could not thereby make this man bald. That conclusion is clearly false. It is highly controversial how we should respond to such *sorites arguments*. These are like Zeno's arguments, in ancient Greece, for the impossibility of motion. These were excellent arguments, which were answered only several centuries later when mathematicians reached a better understanding of infinite sequences. But even before these arguments were answered, the ancient Greeks rightly assumed that these arguments must be unsound. It is clear that some things move. Similar claims may apply to sorites arguments and to other arguments against the possibility of indeterminacy.

If some normative questions are indeterminate, having no answer, this would provide another explanation of some normative disagreements. When people disagree about whether some act is wrong, they may mistakenly assume that this act must either be, or not be, wrong. If these people gave up this assumption, they might often cease to disagree.

Such indeterminacy may also partly solve another problem. Return to the question of how much we rich people ought to give to those who are very poor. Now that each of us can so easily save so many other people from death, disablement, and painful diseases, all plausible moral views require us to give a great deal. These views may seem too demanding. If I am regularly giving substantial amounts to some aid agency, I may think that I am doing well enough. *But I could save some young mother's life, at very little cost to myself. And save another's, and save another's.* We can be knocked over or pulled apart by such thoughts. For most readers of this book, this will be their greatest moral challenge. Most of us will not give enough, and will fail in one of two ways. We may have defensible moral beliefs, but only at the cost of breaking the link between our moral beliefs and our intentions. We must then admit that we intend to act wrongly. Or we may keep this link, intending never to act wrongly, but only at the cost of having indefensible moral beliefs. There is, however, another possibility. If we give to the world's poorest people one hundredth of our income, that is too little, and we are acting wrongly. If we gave nearly everything, that would be enough, and we would not

be acting wrongly. But this question may sometimes have no answer. If we give certain proportions of our income, such as one tenth, or one quarter, it may not be true that we are *not* acting wrongly. But it may also not be true that we *are* acting wrongly.

The Argument from Disagreement is sometimes claimed to have most force when it appeals to history. As Nietzsche writes: "Because our moral philosophers . . . were poorly informed and not even very curious about different peoples, times, and past ages – they never laid eyes on the real problems of morality; for these emerge only when we compare *many* moralities" (Nietzsche ME). It is true that in the more distant past people held moral beliefs that conflict more strongly with our present beliefs. This fact would count against the view that we all have moral knowledge, since everyone's conscience infallibly tells us which acts are wrong. But that view is clearly false. Our claim should be only that in ideal conditions we would nearly all have sufficiently similar moral beliefs. This Convergence Claim is not threatened by the fact that in earlier ages, people held moral beliefs that conflict more strongly with our present beliefs. On the contrary, this fact *supports* this claim. As Nietzsche admits, the earliest known moral concepts and moral codes were primitive and crude. When we look at the history of morality, we do not find mere variation, or a jumble of different moralities. We find a series of challenges to established beliefs, which lead to plausible revisions and to greater agreement.

Some examples are beliefs about the scope of the moral community. In many of the earliest moralities, this community excluded slaves and people in other tribes or cities, and gave a lesser status to serfs, peasants, people in lower castes, and women. There has been slow but accelerating progress toward the beliefs that everyone's well-being matters equally and that everyone has equal moral claims.

I have now described many ways in which, when different people seem to have conflicting normative beliefs, these cases may not involve pure normative disagreements. These people may be considering borderline cases, or they may not know all of the relevant facts, or they may have conflicting nonnormative or metaethical beliefs, or they may not understand the relevant arguments, or they may be using different concepts, or be affected by some distorting influence, or they may fail to realize that many normative truths are matters of degree, or that many of these truths are very imprecise, or that some normative questions may not have answers. We can also plausibly believe that, partly by learning from these disagreements, we are making normative progress. These facts do not show that in ideal conditions we would nearly all have sufficiently similar normative beliefs. But when we consider most actual disagreements, these disagreements do not, I believe, count strongly against this Convergence Claim.

We can next note that when we consider some important questions, we *already* have sufficiently similar normative beliefs. With some fairly trivial exceptions, Williams assumes that we cannot claim to have made moral progress, and that there are no moral truths. In defending this skeptical view, Williams appeals to the fact that there have been deep moral disagreements. At one point, Williams writes:

"No doubt there are some ethical beliefs, universally held and usually vague . . . that we can be sure will survive at the reflective level. But they fall far short of any adequate, still less systematic body of ethical knowledge . . ." (Williams 1985: 148). As Williams himself points out, however, ethical knowledge does not have to be *systematic*. Williams rightly criticizes Sidgwick for making that assumption. It *would* matter if, as Williams claims, the universally held beliefs that survived reflection would not even give us an *adequate* body of ethical knowledge. But that is not, I believe, true.

When Williams concedes that there are some vague, universally held moral beliefs, his example is:

(U) One has to have a special reason to kill someone.

We can make this claim less vague. It has long been almost universally believed that

(V) except in certain special cases, it is wrong to kill any innocent human being who is a member of our moral community.

It is now almost universally believed that

(W) this community at least includes all human beings.

There is some disagreement about which are the special cases in which it is *not* wrong to kill some innocent human being. There are also disagreements about what counts as a living human being; about which human beings are, in the relevant sense, innocent; and about what counts as a killing. But these are all disagreements about borderline cases. Many thousands of innocent people are intentionally killed each year. In nearly all these kinds of case, if everyone knew the relevant facts we would nearly all agree about whether these acts are wrong.[9] There are several other important moral beliefs that are nearly universal. Many people act in ways that we nearly all believe to be wrong, and such acts would be much more common if they were not believed to be wrong. Though these beliefs are vague, and there is disagreement about borderline cases, we can justifiably believe that most of these acts *are* wrong.

The Double Badness of Suffering

There are some other normative beliefs which are *not* vague and on which we have already reached sufficient agreement. Few people have denied that

(A) it is in itself bad to suffer.

All suffering is, in this sense, bad *for the sufferer*. Of those who believe that events can be *impersonally* bad, or bad, *period*, few have denied that

(B) it is bad when people suffer in ways that they do not deserve.

These claims describe what we can call the *double badness* of suffering. Though suffering is always *in itself* bad, some suffering has good effects which may make it on the whole good, as when the pain that is caused by some injury prevents us from acting in ways that would increase this injury.

Some people believe that

(C) suffering is in itself impersonally good, or is at least not in itself bad, when and because this suffering is deserved.

This belief does not conflict with (A), since such suffering is thought to be deserved as a punishment, which it could not be if it was not, at least in one way, bad for the sufferer.

Though some people have seemed to deny the double badness of suffering, these people were either not really denying (A) or (B) or they were under the influence of some distorting factor, or both. The Stoics, for example, wanted to believe both that

(D) everything is for the best,

and that

(E) those who were virtuous and wise would have a kind of happiness that did not depend on luck, or on how these people were treated by others.

These claims could not be true if it is bad to suffer. The Stoics therefore claimed that suffering is not bad, and that a wise and virtuous man would be happy even while he was being tortured on a rack.

Though they made such claims, the Stoics did not really deny that it is bad to suffer. These people distinguished two kinds of badness, or disvalue, to one of which they gave a misleading name. Pain and suffering were called *dispreferred indifferents*. Though these states were called *indifferent* in the sense that they had no disvalue of the more important kind, they were called *dispreferred* in the sense that a wise man would try to avoid these states when such attempts were compatible with virtue. When the Stoics called pain and suffering *dispreferred*, they really meant that these states were *dispreferable*, or nonmorally bad in the reason-implying sense. That is why a wise man would try to avoid these states. As Williams points out, there was another tension in the Stoic view. If pain and suffering are not bad, why is cruelty, as the Stoics claimed, a vice?

Many later thinkers have claimed, mostly as one part of a theistic view, that everything is for the best. On one version of this view, held for example by Albertus Magnus, the concepts *real*, *good*, and *created by God* are quite different, since these concepts are expressed by words with quite different meanings, but these concepts all refer to the same property. This view is a fine precursor of Non-Analytical Naturalism. Of these three concepts, the concept *real* is the one that most clearly refers to a property that we can recognize, and that we know some things to have. When we are in great pain, for example, we know what it is for our painful sensation to be real. We also know that some innocent beings suffer in ways that are undeserved, as is true when a trapped fawn is burnt by some forest fire. This fawn's suffering cannot be bad, Albertus Magnus would have claimed, since this suffering is real and is therefore good. But on this view, when we have claimed that this fawn's suffering is real, we cannot claim that this suffering has the *different* property of being good, since there is no such different property. Since this view denies that there are any such independent normative properties, it does not seriously challenge the belief that undeserved suffering is bad. On a closely related view, the *privation theory*, evil is claimed to be merely the absence of good. Undeserved suffering is bad only in the sense that being in agony is not better than being unconscious.

When people make these implausible claims, they are trying to explain why an omniscient, omnipotent, and wholly good God allows what seem to be pointless evils. If undeserved suffering is bad, it is hard to understand why God allows such suffering to occur. Since these people deny that such suffering is bad because this denial seems to them the only solution to this *problem of evil*, these are not clear cases of undistorted disagreement with the view that suffering is bad. Discussing the many weaknesses and errors in our philosophical and other theoretical beliefs, Hume writes: "two thousand years with such long interruptions and under such mighty discouragements are a small space of time to give any tolerable perfection to the sciences" (Hume *Treatise of Human Nature*: Book I, Part IV, Section VII). It is one such interruption to our moral thinking that, for many centuries, many people have believed that everything must be, in some way, good.

Of those who hold such views, as I have said, some use normative words in unusual and irrelevant senses. Another example is Kant's early defense of Alexander Pope's claim, "Whatever is, is right." This claim was mistranslated into German as "Whatever is, is good." Schneewind writes:

> Kant . . . takes perfection to be the relation between the conscious desire to bring some state of affairs into being and the existence of a state of affairs that fully realizes this desire . . . it is plain that . . . Pope's thesis is true, since whatever is, is as a result of God's willing and so is perfect by definition . . . The problem of physical evil [or the badness of pain] is resolved: there simply is none.
>
> (Schneewind 1998: 496)

Kant's definition does not, however, provide a solution. The problem of evil is in part that

(F) God seems to will the existence of a world in which there are some things that are in themselves bad, such as undeserved suffering.

If Kant claimed that such suffering was good or perfect in his special sense, he would mean only that

(G) undeserved suffering that is willed by God is willed by God.

This claim cannot show that such suffering is not bad. Though Kant elsewhere warns that concealed tautologies are trivial, he forgot that here.

There is another way in which when people deny that suffering or pain is bad, they may not be using words like "bad" in relevant senses. Kant was not doing that, for example, when he later defended the Stoic view that physical pain is not bad. Kant meant only that such pain is not morally bad, in the sense in which people and their acts can be bad. Kant is not denying that physical pain is bad in the nonmoral sense of being a state that we have reasons to want not to be in. Ross uses "bad" in another irrelevant sense. When Ross denies that his own pain is bad, he means only that his pain is not something that he has a prima facie duty to prevent.

There are some other people who seem to deny that suffering is bad. One example is Nietzsche. But as I argue elsewhere, this is not really Nietzsche's view (Parfit 2011: ch. 35). There are also some metaethical skeptics, whose doubts are irrelevant here. I know of no one who has both understood the claim that suffering is doubly bad, in the reason-implying senses, and also in an undistorted and unbiased way rejected this claim. The double badness of suffering is already, I believe, very close to being a universally recognized truth.

Though my examples have involved physical pain, these claims also apply to mental suffering. Such suffering can be much worse than much physical pain. Of those who have never been severely depressed, for example, many do not realize how awful this state of mind can be. And many of those who kill themselves are not trying to avoid physical pain. When we ask which things can be very bad, the only plausible answers are great suffering and morally bad people, mental states, and acts. There are many things that may be in themselves good, but the absence of these things is not in itself bad. Friendship, love, knowledge, and various achievements may be in themselves good, but solitude, ignorance, and inactivity are not in themselves bad. False beliefs have been claimed to be bad, but they could not be, in themselves, great evils. And when Moore claims that it is in itself very bad to enjoy looking at ugly things, this is mere aesthetic snobbery.[10]

Though I have claimed only that

(B) undeserved suffering is in itself impersonally bad,

I believe that

(H) no one could ever deserve to suffer,

so that

(I) all suffering is in itself both bad for the sufferer and impersonally bad.

Unlike (B), however, (H) and (I) are not yet universally recognized truths. And unlike those who believe that everything is for the best, some of those who have rejected (H) have not been obviously affected by some distorting influence. I can only hope that in ideal conditions these people would accept both (H) and (I). There may be some undiscovered argument by which, at last, such people will be convinced.

Notes

1 This material is taken from Parfit (2011: ch. 34).
2 Schiffer (2003: 247–8). In describing his imagined people, Schiffer calls them "equally intelligent, rational, imaginative, and attentive (please feel free to insert whatever I left out)" (243–4). I have inserted that these people are rational in the sense that they would respond to reasons.
3 This may be what Boghossian means, in claiming (G).
4 It might be claimed that (L) is not trivial, because this claim implies that some acts are wrong. But in the sense that makes (L) true by definition, (L) means that if certain acts are wrong we ought not to act in these ways. Moral nihilists could accept (L), but deny that any acts are wrong.
5 Schiffer may be one such person. When Schiffer claims that certain facts give us a reason to have some belief, he may not use the word "reason" in the irreducibly normative sense that I express with the phase "counts in favor." Schiffer says that he doubts whether epistemic reasons are normative.
6 As I argue in Parfit (2011: ch. 2, 24 and 30).
7 I discuss these views further in Parfit (2008).
8 For discussions of both kinds of case, see the articles in Chang (1997).
9 The main exceptions are wars and capital punishment.
10 Moore writes that "aesthetic appreciation" of "what is positively ugly . . . is certainly often positively bad in a high degree" (1903: 239).

References

Boghossian, Paul (2000) *New Essays on the A Priori*, eds. Paul Boghossian and Christopher Peacocke, Oxford University Press.

Boghossian, Paul (2008) *Content and Justification*, Oxford University Press.
Chang, Ruth, ed. (1997) *Incommensurability, Incomparability and Practical Reason*, Harvard University Press.
Field, Hartry (2000) "Apriority as an Evaluative Notion" in *New Essays on the A Priori*, eds. Paul Boghossian and Christopher Peacocke, Oxford University Press, pp. 117–49.
Hume, David *Treatise of Human Nature*, ed. L.A. Selby-Bigge, Oxford: Oxford University Press, Book I, Part IV, Section VII, p. 273.
Moore, G.E. (1903) *Principia Ethica*, Cambridge University Press.
Nietzsche, F. (ME) *Beyond Good and Evil*, various publishers and dates.
Parfit, Derek (2008) "Persons, Bodies, and Human Beings," in *Contemporary Debates in Metaphysics*, ed. John Hawthorne, Dean Zimmerman, and Theodore Sider, Wiley-Blackwell, pp. 177–208.
Parfit, Derek (2011) *On What Matters*, Oxford University Press.
Scanlon, T.M. (2003) "Rawls on Justification," in *The Cambridge Companion to Rawls*, ed. Samuel Freeman, Cambridge University Press, pp. 139–67.
Schiffer, Stephen (2003) *The Things We Mean*, Oxford University Press.
Schneewind, Jerome (1998) *The Invention of Autonomy*, Cambridge University Press.
Sen, Amartya (1973) *On Economic Inequality*, Oxford University Press.
Sidgwick, Henry (ME) *The Methods of Ethics*, Macmillan and Hackett, various dates.
Williams, Bernard (1985) *Ethics and the Limits of Philosophy*, Fontana.

Chapter 4

Divine Command Theory

Philip L. Quinn

Judaism, Christianity, and Islam share the view that the Hebrew Bible has authority in matters of religion. They therefore have reasons for sympathy with a divine command conception of morality. Both Exodus 20:1–17 and Deuteronomy 5:6–21, which recount the revelation of the Decalogue, portray God as instructing the Chosen People about what they are to do and not to do by commanding them. One might, of course, understand these divine commands as merely God's endorsement of a moral code whose authority is independent of the commands. But it seems natural enough to suppose that the authority of the Decalogue depends in some manner on the fact it is divinely commanded or the fact that the commands express God's will. So the major monotheisms have reasons to develop accounts of morality according to which it depends upon God. A long tradition of theological voluntarism in moral theory has evolved from this natural starting point.

During roughly the last quarter of the twentieth century, there has been a revival of interest in divine command morality within the community of analytic philosophers of religion. Attention has been paid to three important questions: How can the idea that morality depends upon God best be spelled out and given a precise theoretical formulation? How can the theory thus formulated be supported by argument? And how can that theory be defended against objections? In the three sections of this essay, I propose answers to these questions.

The Blackwell Guide to Ethical Theory, Second Edition. Edited by Hugh LaFollette and Ingmar Persson.

Formulating the Theory

Settling on a precise theoretical formulation of the idea that morality depends upon God involves addressing three issues. The following schema can be used to indicate what they are:

(S) Moral status M stands in dependency relation D to divine act A.

The first issue is the specification of the moral statuses that the theory will claim are dependent on God. The second is specifying the nature of the dependency relation the theory will assert holds between God and those moral statuses. And the third is specifying the divine acts on which the moral statuses will be said by the theory to depend. Each of the three specifications involves a choice among options.

There is general agreement that the theory should claim that some or all the deontological moral statuses depend upon God. Those statuses are moral requirement (obligation), moral permission (rightness), and moral prohibition (wrongness). This agreement is understandable if one thinks of God's will or commands as creating moral law, for then the deontological moral statuses are analogous to the ordinary categories of legal requirement, permission, and prohibition. I once proposed a theory according to which the axiological statuses of moral goodness, moral badness, and moral indifference also depend upon God (Quinn 1978: 67–73). Other theorists, however, have restricted their attention to the deontological moral statuses. In the present discussion, I will follow their lead and formulate a theory in which only deontological moral statuses depend upon God.

Several accounts of the dependency relation have been proposed in recent years. In the seminal paper that started the revival of interest in divine command morality, Robert M. Adams (1973) proposed a theory in which being contrary to the commands of a loving God is part of the meaning of being morally wrong in the discourse of some Jewish and Christian believers. My initial proposal in Quinn (1978) was that divine commands and moral requirements are necessarily coextensive. Adams (1979) records a switch to the view that the property of moral wrongness is identical to the property of being contrary to the commands of a loving God.

I think all these proposals are flawed in one way or another. There may be Jews and Christians in whose discourse being morally wrong means, in part, being contrary to the commands of a loving God. But this is not the case in my discourse, and so the initial proposal made by Adams does not present a divine command theory I could accept. The flaw in my initial proposal is that it fails to capture the asymmetry of the dependence of moral requirements on divine commands because necessary coextensiveness is a symmetrical relation. And it seems clear to me that being morally wrong and being contrary to the commands of a loving God are distinct properties. Perhaps being morally wrong supervenes on being contrary to

the commands of a loving God. Since supervenience relations are supposed to be asymmetrical, this possibility is worth exploring, though, as far as I know, no one has formulated a divine command theory in supervenience terms.

My second proposal in Quinn (1979) was a theory in which divine commands are necessary and sufficient causal conditions for moral requirements. Edward R. Wierenga (1989) has formulated a theory in which, by commanding a person to bring about a state of affairs at a time, God brings it about that it is obligatory that the person bring about the state of affairs at the time. My current view is that dependence of morality on God is best formulated in terms of a relation of bringing about, though care must be taken to distinguish this relation from various causal relations familiar from science and ordinary life. In particular, the divine bringing about in question will have the following marks: totality, exclusivity, activity, immediacy, and necessity. By totality, I mean that what does the bringing about is the total cause of what is brought about. By exclusivity, I mean that what does the bringing about is the sole cause of what is brought about. By activity, I mean that what does the bringing about does so in virtue of the exercise of some active power. By immediacy, I mean that what does the bringing about causes what is brought about immediately rather than by means of secondary causes or instruments. And by necessity, I mean that what does the bringing about necessitates what is brought about.

There is controversy about which divine acts bring about moral requirements, permissions, and prohibitions. As I see it, it is at the deepest level God's will, and not divine commands – which merely express or reveal God's will – that determines the deontological status of human actions. But Adams (1996) has recently objected to replacing divine commands with God's will in formulating the theory. It is therefore incumbent on me to respond to his objections. There are two of them.

The first derives from a problem about how to think about God's will. Theologians often distinguish between God's antecedent will and God's consequent will. As Adams understands the distinction, "God's antecedent will is God's preference regarding a particular issue considered rather narrowly in itself, other things being equal. God's consequent will is God's preference regarding the matter, all things considered" (Adams 1996: 60–1). It is commonly held that nothing happens contrary to God's consequent will, which is partly permissive. But since wrong actions do occur, wrongness cannot be specified in terms of contrariety to God's consequent will. Nor, according to Adams, can the ground of obligation be identified with God's antecedent will because we are sometimes morally obliged to make the best of a bad situation by doing something that a good God would not antecedently have preferred, other things being equal. And if we identify the ground of obligation with God's revealed will, we are in effect identifying it with divine commands.

My response to this objection is to deny that divine antecedent preferences, other things being equal, exhaust God's antecedent will. Following a suggestion by Mark Murphy (1998), I also attribute to God's antecedent will intentions, and I think divine antecedent intentions can be used to account for obligations

to make the best of bad situations. Suppose I make a promise. God surely prefers that I keep it, other things being equal. Assume God also antecedently intends that I keep my promise, which makes it obligatory for me to keep it. If I break my promise, I create a bad situation by violating an obligation. But assume that God, in addition, antecedently intends that I apologize if I break my promise, which makes it obligatory for me to apologize if I break it. If I am in the bad situation of having broken my promise, then my obligation is to apologize. If I fail to apologize, I violate a second obligation. Of course, if I both break my promise and fail to apologize, then God neither consequently intends that I keep my promise nor consequently intends that I apologize, for nothing happens contrary to God's consequent intentions. My conclusion is that a sufficiently rich account of God's antecedent will allows us to identify the ground of obligation with some of its activities.

But Adams has a second objection. Replacing divine commands with the divine will as the ground of obligation makes sense only on the assumption that God's will can be what it is without being revealed. According to Adams, this has three undesirable consequences. First, it yields an unattractive picture of divine–human relations "in which the wish of God's heart imposes binding obligations without even being communicated, much less issuing in a command" (Adams 1996: 61). Second, "basing obligation on unrevealed as distinct from revealed divine will deprives God of the freedom to choose whether or not to impose an obligation" (Adams 1996: 61). And third, "it closes off the possibility of supererogation, important in some theistic ethical theories, the possibility of an action that is preferable from God's point of view but not ethically required" (Adams 1996: 61–2).

However, even if divine antecedent intentions impose binding obligations without being communicated, we can be confident that a good God will communicate many of these divine antecedent intentions by means of commands. As Murphy points out, God need not communicate all of them because we can infer some of them from the express divine commands, using principles of rational intending. Moreover, even if God's preferences, other things being equal, or the wishes of God's heart are not freely chosen, there is no reason to believe that God lacks freedom of choice with respect to the formation of the antecedent intentions that impose obligations. And it seems possible for there to be an action that is preferable from God's point of view whose performance is not antecedently intended by God. Hence I do not think placing the ground of obligation in the divine antecedent will has undesirable consequences if we operate with a sufficiently nuanced conception of God's antecedent will.

Thinking both of the objections set forth by Adams can be answered, I stick with my intuition that God's will determines the deontological status of actions. Revising some principles found in Wierenga (1989: 216–17) to reflect this option, I propose that the best theoretical formulations of the idea that the deontological part of morality depends upon God consists of the following three principles:

(P1) For every human agent *x*, state of affairs *S*, and time *t*, (1) it is morally obligatory that *x* bring about *S* at *t* if and only if God antecedently intends that *x* bring about *S* at *t*, and (2) if it is morally obligatory that *x* bring about *S* at *t*, then by antecedently intending that *x* bring about *S* at *t* God brings it about that it is morally obligatory that *x* bring about *S* at *t*;

(P2) For every human agent *x*, state of affairs *S*, and time *t*, (1) it is morally permissible that *x* bring about *S* at *t* if and only if God refrains from antecedently intending that *x* not bring about *S* at *t*, and (2) if it is morally permissible that *x* bring about *S* at *t*, then by refraining from antecedently intending that *x* not bring about *S* at *t* God brings it about that it is morally permissible that *x* bring about *S* at *t*;

(P3) For every human agent *x*, state of affairs *S*, and time *t*, (1) it is morally wrong that *x* bring about *S* at *t* if and only if God antecedently intends that *x* not bring about *S* at *t*, and (2) if it is morally wrong that *x* bring about *S* at *t*, then by antecedently intending that *x* not bring about *S* at *t* God brings it about that it is morally wrong that *x* bring about *S* at *t*.

Of course, this theory is not, strictly speaking, a divine command theory; it is instead a divine intention theory. It is, however, a version of theological voluntarism, and it pictures divine commands as expressing or revealing God's antecedent intentions. So when we speak loosely, I suppose no harm is done if we conduct the discussion in terms of divine commands. In what follows, I will do this, occasionally reminding the reader that it is the divine intentions lying behind the divine commands that really make a moral difference.

Supporting the Theory

I know of no deductive argument that is a proof of the theory I have formulated or of any of its near neighbors. In Quinn (1990a), I constructed a valid deductive argument, from premises widely acceptable among theists, for the conclusion that many obtaining deontological states of affairs are metaphysically dependent on the will of God; but I noted that this argument could not provide good reasons for holding that all such states of affairs are so dependent. I am now inclined to doubt that constructing deductive arguments is the most promising way of supporting theological voluntarism. I think a more fruitful approach is to support it by a cumulative case argument. In Quinn (1990b) and Quinn (1992), I began to construct a cumulative case for theological voluntarism. In the present discussion, I will summarize and extend my previous arguments. My cumulative case now has four parts. They support theological voluntarism in ways analogous to that in which the legs of a chair support the weight of a seated person. No one leg

supports all the weight, but each leg contributes to supporting the weight. I do not claim that my cumulative case for theological voluntarism is a complete case or the strongest case that could be made. I think all parts of my cumulative case should have some attractiveness for Christians. One of its parts will appeal only to Christians; two others may appeal to both Christians and some other theists; and the final part should appeal to all monotheists. I do not expect my cumulative case to persuade any nontheists to become theological voluntarists; however, I hope it will convince some nontheists that theological voluntarism is an attractive option for theists. I begin with the part with narrowest appeal and end with the part with broadest appeal.

Commanded Christian Love

It is a striking feature of the ethics of love set forth in the New Testament that love is commanded. In Matthew's Gospel, Jesus states the command in response to a question from a lawyer about which commandment of the law is the greatest. He says: "You shall love the Lord your God with your whole heart, with your whole soul, and with all your mind. This is the greatest and first commandment. The second is like it: You shall love your neighbor as yourself" (Matthew 22:37–9). Mark 12:29–31 tells of Jesus giving essentially the same answer to a scribe, and Luke 10:27–8 speaks of a lawyer giving this answer to a question from Jesus and being told by Jesus that it is correct. In his last discourse, recorded in John's Gospel, Jesus tells his followers that "the command I give you is this, that you love one another" (John 15:17). So the authors of these books concur that the Christian ethics of love for one another is expressed in the form of a command. If Jesus is God the Son, this command and the intention behind it are divine.

Is there a reason for love of neighbor being made a matter of obligation or duty? I think there is. It is that the love of neighbor of which Jesus speaks is extremely difficult for humans in their present condition. It does not spontaneously engage their affections, and so, if it were merely permissible, they would not love their neighbors. It is therefore no accident that the love of neighbor Jesus endorses is a commanded love.

In my view, no one has seen with greater clarity than Kierkegaard just how radical the demands of love of neighbor are. In *Works of Love*, his discourse on Matthew 22:39 draws a sharp distinction between erotic love and friendship, on the one hand, and Christian love of neighbor, on the other. Both erotic love and friendship play favorites; the love of neighbor Christians are commanded to display is completely impartial. Kierkegaard says: "The object of both erotic love and friendship has therefore also the favorite's name, *the beloved*, *the friend*, who is loved in distinction from the rest of the world. On the other hand, the Christian teaching is to love one's neighbor, to love all mankind, all men, even enemies, and not to make exceptions, neither in favoritism nor in aversion" (Kierkegaard 1847/1964: 36). His shocking idea is that the obligation to love imposed by the

command places absolutely every human, including one's beloved, one's friend, and one's very self, on the same footing as one's worst enemy or millions of people with whom one has had no contact. Perhaps it is easy to imagine God loving all humans in this undiscriminating way. It is hard to see how it could be either desirable or feasible for humans to respond to one another in this fashion. But if Kierkegaard is right, this is exactly what the command to love the neighbor obliges us to do.

According to Kierkegaard, there is another way in which love of neighbor differs from erotic love and friendship. Erotic love and friendship depend on characteristics of the beloved and the friend that are mutable. If the beloved loses the traits that made him or her erotically attractive, then erotic love dies. If the friend who was prized for having a virtuous character turns vicious, then the friendship is not likely to survive if one remains virtuous. Love of neighbor, however, is supposed to be invulnerable to changes in its object. Kierkegaard puts the point this way: "No change, however, can take your neighbor from you, for it is not your neighbor who holds you fast – it is your love which holds your neighbor fast. If your love for your neighbor remains unchanged, then your neighbor also remains unchanged just by being" (Kierkegaard 1847/1964: 76). If there is to be such a love that alters not where it alteration finds, it cannot depend on mutable features of the neighbor and ways in which they engage our spontaneous affections and natural preferences. For Kierkegaard, it can have the independence it needs only if it is obligatory, for only then can it be motivated by a stable sense of duty rather than by changeable affections or preferences. In this way, he says, "the 'You shall' makes love free in blessed independence; such a love stands and does not fall with variations in the object of love; it stands and falls with eternity's law, but therefore it never falls" (Kierkegaard 1847/1964: 53). We are obliged to obey eternity's law.

Kierkegaard thus has two reasons for thinking that Christian love of neighbor has to be a matter of obligation. The first is that only a love which is obligatory can be sufficiently extensive in scope to embrace everyone without distinction. Erotic love and friendship are always discriminating, partial, and exclusive. The second reason is that only a love which is obligatory can be invulnerable to alterations in its objects. Erotic love and friendship change in response to changes in the valued features of their objects. As I have argued in more detail in Quinn (1996), they are powerful reasons.

My view is that this commanded love is foundational for Christian ethics; it is also what sets Christian ethics apart from rival secular moralities. The stringency of the obligation to love is likely to give offense. In that respect, it resembles the requirements of impartial benevolence or utility maximization in secular moral theories, which are criticized for setting standards impossibly high or not leaving room for personal projects. Kierkegaard wants his readers to see just how demanding the obligation is and to accept it as binding them. "Only acknowledge it," he exhorts them, "or if it is disturbing to you to have it put in this way, I will admit that many times it has thrust me back and that I am yet very far from the illusion that I fulfill this command, which to flesh and blood is offence, and to

wisdom foolishness" (Kierkegaard 1847/1964: 71). I concur with Kierkegaard about the importance of highlighting rather than downplaying the stringency of the obligation to love the neighbor even if, as a result, many people are thrust back or offended. Christians who believe that humans in their present condition are fallen should not find this response surprising. It is only to be expected that people in such a condition will feel comfortable with moral laxity and be offended by moral stringency. There is, however, no reason for Christians to believe that fallen humans have no obligations whose stringency makes them uncomfortable. Loving everyone as we love ourselves is, I think, obligatory in Christian ethics, and it has that status, as the Gospels show us, because of God. It seems to me that Christians who take the Gospels seriously are not in a position to deny that they teach us that God intends us to love the neighbor and has commanded us to do so or that these facts place us under an obligation to love the neighbor. So I find, in what is most distinctive about the Christian ethics of love in the Gospels, a reason for Christians to favor a divine command conception of moral obligation.

Lex Orandi, Lex Credendi

According to an old saying, the law of prayer is the law of belief. The old saying probably should not be regarded as an exceptionless generalization, for popular devotion sometimes contains superstitious elements. But often enough in Christianity what is professed in religious practice is a good guide to what ought to be contained in sound religious theory. Janine M. Idziak (1997) has shown that Christian practice emphasizes the theme of conformity to the divine will.

This theme occurs in classics of Christian spirituality and in the thought of Christian saints. In the late medieval treatise *The Imitation of Christ*, Thomas à Kempis portrays Christ as counseling a disciple "to learn perfect self-surrender, and to accept My will without argument or complaint" (quoted in Idziak 1997: 457). The colonial American saint Elizabeth Seton says that "the first purpose of our daily work is to do the will of God; secondly, to do it in the manner he wills; and thirdly, to do it because it is his will" (quoted in Idziak 1997: 457).

Nor is the theme of conformity to God's will characteristic of only the thought of extraordinary Christians; it is also found in traditional hymns: "Father, who didst fashion man / Godlike in thy loving plan / Fill us with that love divine / And conform our wills to thine" (quoted in Idziak 1997: 457). And it is found in books of worship such as the Presbyterian *Daily Prayer*: "Eternal God, send your Holy Spirit into our hearts, to direct and rule us according to your will . . ."; "God of love, as you have given your life to us, so may we live according to your holy will revealed in Jesus Christ . . ." (quoted in Idziak 1997: 457). These examples, and others like them, make it clear that conformity with the divine will is an important theme in Christian spirituality. Theological voluntarism in ethics expresses this theme at the level of moral theory. There seems to me to be nothing

superstitious in this aspect of Christian practice. So I regard it as providing some support for theological voluntarism in moral theory in accord with the principle of *lex orandi*, *lex credendi*. In other words, the fact that conformity to the will of God is an important theme in Christian devotional and liturgical practices is a good reason for Christians to adopt a moral theory in which the will of God is a source of obligations.

I am no expert on the religious practices of Judaism and Islam. What I know about Jewish and Islamic religious thought suggests that Jews and Muslims regard conformity to the will of Yahweh and the will of Allah, respectively, as very important. So I tend to think Jews and Muslims have available to them arguments similar to my argument in the case of Christianity. If so, arguments of this kind will have some appeal not just for Christians but also for some adherents of the other two major monotheisms.

The Immoralities of the Patriarchs

A Christian tradition of interpreting some stories in the Hebrew Bible serves as the basis for an argument to the conclusion that the deontological status of at least some actions depends upon God. These stories recount the incidents sometimes described as the immoralities of the patriarchs. They are cases in which God commands something that appears to be immoral and, indeed, to violate a prohibition God lays down in the Decalogue. Three such cases come up over and over again in medieval discussions. The first is the divine command to Abraham, recorded in Genesis 22:1–2, to sacrifice his son Isaac. The second is the divine command reported in Exodus 11:2, which was taken to be a command that the Israelites plunder the Egyptians. And the third is the divine command to the prophet Hosea, stated first in Hosea 1:2 and then repeated in Hosea 3:1, to have sexual relations with an adulteress. According to these stories, God has apparently commanded homicide, theft and adultery (or at least fornication) in particular cases, and such actions are apparently contrary to the prohibitions of the Decalogue. What should the patriarchs do? How are we to interpret these stories?

The tradition of biblical exegesis I am going to discuss takes the stories to be literally true; it presupposes that God actually did command as the stories say God did. It also assumes that these commands were binding on those to whom they were addressed. In *The City of God*, Augustine uses the case of Abraham to make the point that the divine law prohibiting killing allows exceptions "when God authorizes killing by a general law or when He gives an explicit commission to an individual for a limited time." Abraham, he says, "was not only free from the guilt of criminal cruelty, but even commended for his piety, when he consented to sacrifice his son, not, indeed, with criminal intent but in obedience to God" (Augustine of Hippo 426/1958: bk 1, ch. 21). Augustine thinks God explicitly commissioned Abraham to kill Isaac and then revoked the commission just before

the killing was to have taken place. It is clear that Augustine believes Abraham did what he should do in consenting to kill Isaac because the killing had been commanded by God. He also believes that Abraham's consent, which would have been wrong in the absence of the command, was not wrong given its presence. So Augustine holds that divine commands addressed to particular individuals (or the divine intentions they express) determine the deontological status of actions those individuals perform in obedience to them.

The connection of these cases to divine command ethics is made explicit in the work of Andrew of Neufchateau, a fourteenth-century Franciscan who is judged by Idziak to have conducted "the lengthiest and most sophisticated defense of the position" (Idziak 1989: 63). Andrew claims that there are actions which, "known per se by the law of nature and by the dictate of natural reason, appear to be prohibited, actions such as homicides, thefts, adulteries, etc. But it is possible that such actions not be sins with respect to the absolute power of God" (Andrew 1514/1997: 91). Abraham, he goes on to say, "wished to kill his son so that he would be obedient to God commanding this, and he would not have sinned in doing this if God should not have withdrawn his command" (Andrew 1514/1997: 91). For Andrew, not only did Abraham do no wrong in consenting to kill Isaac but he would have done no wrong if the command had not been withdrawn and he had killed Isaac. In his view, God's absolute power is such that acts such as homicides, thefts, and adulteries, which are seen to be prohibited and so sins when known by means of natural law and natural reason, would not be sins and so would not be wrong if they were commanded by God, as some in fact have been. He shares with Augustine the view that divine commands (or the divine intentions they express) can and do determine the deontological status of actions.

Thomas Aquinas too shares this view. He treats the three cases in the following passage, which deserves to be quoted in full:

> Consequently when the children of Israel, by God's command, took away the spoils of the Egyptians, this was not theft; since it was due to them by the sentence of God. Likewise when Abraham consented to slay his son, he did not consent to murder, because his son was due to be slain by the command of God, Who is Lord of life and death: for He it is Who inflicts the punishment of death on all men, both godly and ungodly, on account of the sin of our first parent, and if a man be the executor of that sentence by Divine authority, he will be no murderer any more than God would be. Again Osee, by taking unto himself a wife of fornications, or an adulterous woman, was not guilty either of adultery or of fornication: because he took unto himself one who was his by command of God, Who is the author of the institution of marriage
>
> (Aquinas 1273/1948: I–II, q.100, a.8, ad 3).

Aquinas reasons in the following manner. Because God commanded the Israelites to plunder the Egyptians, what the Israelites took was due to them and not to the

Egyptians. Since theft involves taking what is not one's due, the plunder of the Egyptians was not theft. Similarly, because God, who is lord of life and death, commanded Abraham to slay Isaac, Isaac was due to receive the punishment of death all humans deserve in consequence of original sin. Since murder involves slaying someone who is not due to be slain, the slaying of Isaac would not have been murder. And because God, who is the author of marriage, commanded Hosea to take the adulteress as his wife, she was his wife, and so he was guilty of neither adultery nor fornication in having intercourse with her.

Andrew and Aquinas differ in some respects about what the divine commands do. Andrew seems to think that God's command to Abraham brings it about that the slaying of Isaac would not be wrong while remaining a murder. Aquinas clearly supposes that God's command to Abraham brings it about that the slaying of Isaac would be neither wrong nor a murder. But such disagreement should not blind one to the ways in which they agree. Both hold that the slaying of Isaac by Abraham, which would be wrong in the absence of the divine command, will not be wrong in its presence if Abraham obeys it. We might sum up the agreement by saying that what divine commands do is to make obligatory patriarchal actions that would have been wrong in their absence. And because divine commands do this in virtue of something necessarily restricted to God alone, such as absolute power or lordship over life and death, human commands could not make a moral difference of this sort.

It is worth noting that agreement with Augustine, Andrew, and Aquinas about such cases need not be restricted to Christians who share their belief that there actually were the divine commands reported in the scriptural stories. Some may choose to think of such cases as merely possible but concur with the tradition of exegesis I have been describing in believing that divine commands would make a moral difference of the sort our medieval interpreters thought they in fact did make. I think there would be enough agreement about such cases among reflective Christians to make it fair to claim that Christian moral intuitions about scriptural cases support the conclusion that God is a source of moral obligation. What is more, it appears to be only a contingent fact that there are at most a few such cases. The properties, such as absolute power or lordship over life and death, in virtue of which divine commands have their moral effects, would still be possessed by God even if such commands were more numerous. So it is hard to resist the conclusion that any act of homicide, plunder, or intercourse with a person other than one's spouse would be obligatory if it were divinely commanded. Thus the intuitions underlying this tradition of exegesis also support the conclusion that whether any action is morally obligatory or not depends on whether it is divinely commanded (or divinely intended) or not.

I cannot speak with authority about how the exegetical traditions of Judaism and Islam treat the incidents known as the immoralities of the patriarchs. It does seem to me, however, that Jews and Muslims have available to them the strategy of interpretation made use of by Augustine, Andrew, and Aquinas. Those among

them who adopt this strategy will be able to use scriptural cases to support the view that Yahweh or Allah is a source of moral obligation.

Absolute Divine Sovereignty

There are several reasons why theists of all stripes – Jews, Christians, and Muslims alike – would favor including a strong doctrine of divine sovereignty in their philosophical theology. Two of the most important pertain to creation and providence. Theists customarily wish to insist on a sharp distinction between God and creation. According to traditional accounts of creation and conservation, each contingent thing depends on God's power for its existence whenever it exists. God, by contrast, depends on nothing external for existence. So God has complete sovereignty over the realm of contingent existence. Theists also usually wish to maintain that we can trust God's eschatological promises without any reservation. Even if God does not control the finest details of history because God has chosen to create a world in which there is microphysical chance or libertarian freedom, God has the power to ensure that the created cosmos will serve God's purposes for it and all its inhabitants in the long run. So God also has extensive sovereignty over the realm of contingent events. Considerations of theoretical unity then make it attractive to extend the scope of divine sovereignty from the realm of fact into the realm of value. It is an extension of this sort that we find in the remark by Andrew of Neufchateau that, with respect to God's absolute power, it is possible for homicides, thefts, and adulteries not to be sins. More controversially, the same considerations make it tempting to extend the scope of divine sovereignty from the realm of the contingent into the realm of the necessary.

How far can such extensions be pushed? In recent work in philosophical theology, Thomas V. Morris (1987) has argued for a view of absolute creation according to which God is the creator of necessary as well as contingent reality. As he sees it, in order to be absolute creator, God must be responsible somehow for the necessary truth of all propositions that are necessarily true. If this view is tenable, he notes, "moral truths can be objective, unalterable, and necessary, and yet still dependent on God" (Morris 1987: 171). Thus, for example, even if it is necessarily true that murder, theft, and adultery are morally wrong, God is responsible, according to the absolute creationist, for the necessary truth of the proposition that murder, theft, and adultery are morally wrong. But how could God be responsible for the necessary truth of a proposition?

Michael J. Loux (1986) has made an interesting suggestion about how necessary truths might depend upon God. It involves the idea that there is an asymmetrical relation of metaphysical dependence between certain divine beliefs and facts being necessarily as they are. Taking notions of believing and entertaining as primitives, Loux defines a concept of strong belief as follows: a person x strongly believes that p if and only if x believes that p and does not entertain that not-p. Since God is omniscient, divine beliefs correlate perfectly with truth and divine

strong beliefs correlate perfectly with necessary truth. But there is more than mere correlation between divine strong beliefs and necessary truth. According to Loux, "God is not in the relevant strong belief states because the facts are necessarily as they are. On the contrary, the facts are necessarily as they are because God has the relevant strong beliefs. So it is the case that 2 + 2 = 4 because God believes that 2 + 2 = 4; and it is necessarily the case that 2 + 2 = 4 because God strongly believes that 2 + 2 = 4" (Loux 1986: 510). And, of course, this idea can be extended to the moral realm. It is the case that murder, theft, and adultery are wrong, on this view, because God believes that murder, theft, and adultery are wrong; and if it is necessarily the case that murder, theft, and adultery are wrong, this is so because God strongly believes that murder, theft, and adultery are wrong.

Loux's idea can also be modified to fit the sort of theological voluntarism previously discussed. Suppose that divine strong antecedent intentions are antecedent intentions that God, being essentially perfectly good, could not have failed to form. According to our theory, it is the case that murder, theft, and adultery are morally wrong because God antecedently intends that no one ever bring about the state of affairs of an act of murder, theft, or adultery being performed. It is by antecedently intending that no one ever bring about the state of affairs of an act of murder, theft, or adultery being performed, according to (P3), that God brings it about that it is the case that murder, theft, and adultery are morally wrong. The extension into the realm of the necessary is straightforward. If it is necessarily the case that murder, theft, and adultery are morally wrong, this is so because God strongly antecedently intends that no one ever bring about the state of affairs of an act of murder, theft, or adultery being performed. It is by strongly antecedently intending that no one ever bring about the state of affairs of an act of murder, theft, or adultery being performed, according to the natural extension of (P3), that God brings it about that it is necessarily the case that murder, theft, and adultery are morally wrong. And (P1) and (P2) can be extended in similar ways.

Less formally but more generally, the idea is that moral facts about deontological status are as they are because God has certain antecedent intentions concerning the actions of creaturely moral agents, and necessary moral facts about deontological status, if there are any, are as they are because God has certain strong antecedent intentions concerning the actions of creaturely moral agents. This idea gets support from the doctrine of divine sovereignty because it extends God's sovereignty to cover both the contingent part and, if there is one, the necessary part of the deontological realm.

I think the strength of my cumulative case for theological voluntarism derives in part from the diversity of sources to which it appeals. The ethical demands set forth by Jesus in the Gospels, considerations drawn from religious practice, commentary on incidents portrayed in the Hebrew Bible, and considerations from philosophical theology converge in supporting the position. Further support may be available from arguments to be found in medieval and early modern discussions of divine command ethics. Idziak (1989) contains a catalogue of such arguments.

Perhaps some of these arguments can be updated and made parts of a contemporary cumulative case for theological voluntarism.

Defending the Theory

Before the recent revival of interest in divine command theory began, many philosophers were convinced that objections sufficient to refute theological voluntarism were known. So, particularly during the earlier phases of the revival, a lot of energy went into defending the theory against objections. A successful defense shows that the objections fail to establish the falsity of the theory. Each objection must be considered on its own merits, and objections must be replied to one by one. I have replied to a total of fourteen such objections in Quinn (1978) and Quinn (1979). Adams (1973) and Wierenga (1989) have replied to others. Richard J. Mouw (1990) has replied to yet others. I do not have space in this essay even to summarize all these objections and replies. I will, however, present and reply to five objections. They are the ones that, in my experience, many people find most troubling.

The Trivial Natural Theory Objection

According to this objection, if we accept theological voluntarism, the task of proving the existence God is trivialized. Consider the following argument.

(1) It is morally wrong for some human agents to bring about some states of affairs at some times.
(2) For all human agents, states of affairs, and times, if it is morally wrong for an agent to bring about a state of affairs at a time, then God antecedently intends that the agent not bring about the state of affairs at the time.
(3) Hence, God antecedently intends that some human agents not bring about some states of affairs at some times.
(4) Therefore, God exists.

This argument is deductively valid. Almost everyone will admit the truth of its first premise. It would be hard to defend the claim that there are no wrong actions. The second premise is a consequence of (P3), our theory's principle of wrongness. So if there are wrong actions and our theory is true, the argument is sound. It thus looks as if the theological voluntarist has a simple way to prove the existence of God. However, it does not seem reasonable to believe that this argument proves the existence of God. Nor does it seem reasonable to deny that there are morally wrong actions. So it must be unreasonable to believe that the argument's second

premise is true. And therefore it must be unreasonable to believe that our theory, from which that premise follows, is true.

In response to the objection, a theological voluntarist can point out that not every sound argument is a successful proof of its conclusion. Consider, for example, this argument:

(5) There are humans.
(6) Either there are no humans or God exists.
(7) Hence, God exists.

As in the previous case, the argument is deductively valid, and almost everyone will admit the truth of its first premise. If one then acknowledges the truth of its second premise, one is committed to admitting its soundness and the truth of its conclusion. But even a theist, who will, of course, believe that this argument is sound, need not concede that it is a successful proof of the existence of God. It is not easy to say precisely why the argument is not a successful proof. Perhaps the reason is that, even for a theist, the second premise cannot have, apart from the argument, more epistemic justification than the conclusion, in which case the argument cannot, as a successful proof must, transmit epistemic justification from its premises to its conclusion.

It is open to the theological voluntarist to say similar things about the previous argument. The theological voluntarist will believe that it is sound but need not consider it a successful proof of the existence of God. The reason the argument is not a successful proof, it can be claimed, is that its second premise is not, apart from the argument, better justified epistemically, even for the theological voluntarist, than its conclusion. We would not expect anyone who believes that there are wrong actions to think that theological voluntarism has more epistemic justification than theism itself. It does not follow, as the objection has it, that it is unreasonable to believe that the argument's second premise is true. The lesson to be learned is that the argument's second premise cannot antecedently be more reasonable than its conclusion for anyone who holds that there are wrong actions. Hence the argument cannot, as a proof must, make it more reasonable to believe its conclusion than it otherwise would be. More generally, theological voluntarism does not contribute to the epistemic justification of theism; the order of epistemic justification goes in the other direction.

It is worth noting that sound arguments which do not prove their conclusions are ubiquitous. Here is one:

(8) Horses exist.
(9) Either there are no horses or oysters exist.
(10) Hence, oysters exist.

This argument is obviously sound, but it is no proof of the existence of oysters. Presumably the reason it is not a proof is that, given that we know (8), our

justification for believing (9) is our independent knowledge of (10). Such examples show that the theological voluntarist's response to the objection does not rely on an esoteric feature of argument in natural theology.

The Moral Skepticism Objection

It is sometimes thought that theological voluntarism inevitably leads to moral skepticism. An argument in support of this view might go along the following lines. According to theological voluntarism, we can come to know what is morally obligatory, permissible, and wrong only by first coming to know certain facts about the divine will. But we cannot, at least in this life, come to know such facts, for God's will is inscrutable. Hence, we cannot in this life come to know what is morally obligatory, permissible, and wrong. A more modest version of this objection is the complaint that, according to theological voluntarism, only people who have religious knowledge can have moral knowledge. As Eric D'Arcy puts it, "If immoral actions are immoral merely because God so wills it, merely because God legislates against them, it would be sheer coincidence if someone who knew nothing of God or his law happened to adopt the same views about particular actions as God did" (D'Arcy 1973: 194). And, of course, mere coincidence of our views with God's views, though it would give us true beliefs, would not suffice for moral knowledge.

One reply to the objection is to deny that the divine will is inscrutable. The theological voluntarist can appeal to scripture, religious tradition, personal revelation, and even natural law as sources of knowledge concerning what God has willed. But then the skeptical worry will shift to the disagreements among religious people about what the deliverances of those sources are. Another reply gets closer to the heart of the matter. Our theory asserts that divine antecedent intentions bring it about that certain things are morally obligatory, others are permissible, and others are wrong. It makes no claims in moral epistemology, and so it makes no claims about how we might come to know what God's antecedent intentions are. It does not entail that we can come to know what is morally obligatory, permissible, and wrong only by first coming to know what God's antecedent intentions are. It is consistent with the view that we can only come to know what God's antecedent intentions are by first coming to know what is morally obligatory, permissible, and wrong. This is as it should be. The subject matter of our theory is a certain kind of metaphysical dependency of deontological status on divine intentions. The order of epistemic access may run in the opposite direction from the order of metaphysical dependency. After all, though effects are metaphysically dependent on their causes, in ordinary life we often come to know causes by first coming to know their effects. It is not a consequence of our theory that only people who have religious knowledge can have moral knowledge. Hence, the objection fails.

Whether or not agreement of the views of those who know nothing of God with God's views about the morality of actions is mere coincidence depends on the explanation of the agreement. An explanation available to theological voluntarists is that God has benevolently endowed normal human creatures with a moral faculty such as conscience that, when functioning properly in appropriate circumstances, reliably tracks, unbeknownst to those who know nothing of God, divine antecedent intentions. If that explanation is correct, the agreement is not mere coincidence, and, on reliabilist accounts of knowledge, those who know nothing of God are not precluded from having moral knowledge.

The Uselessness Objection

It is sometimes argued that theological voluntarism is useless as an ethical standard. Jeremy Bentham says: "We may be perfectly sure, indeed, that whatever is right is conformable to the will of God: but so far is that from answering the purpose of showing us what is right, that it is necessary to know first whether a thing is right, in order to know from thence whether it be conformable to the will of God" (Bentham 1789/1948: 22). So his view is that we can come to know what is conformable to the divine will only by first coming to know what is right. Many theological voluntarists would disagree with this view and argue that sometimes we can come to know what is conformable to the will of God from such sources as revelation. But Bentham's view is consistent with our theory. If it is correct, our theory does not provide a decision procedure for the deontological part of ethics: a way of deciding or determining what is right. However, our theory makes no claim to provide a decision procedure. Ethical theories can perform functions other than teaching us how to decide what is right. It would be of theoretical interest to find out that what is morally obligatory, permissible, and wrong depends on divine antecedent intentions, even if this knowledge were not of any practical use. So even if Bentham's view were correct, it would not constitute a successful objection to our theory. Moreover, it is worth noting, by way of an ad hominem against Bentham, that his brand of utilitarianism would be in trouble if this objection were cogent. No one is in a position to calculate the exact hedonic values of all the consequences of all the alternative actions open to an agent in many circumstances in which moral decisions must be made. Nonetheless, a utilitarian may reply, it would be of theoretical interest to find out that hedonistic act-utilitarianism is true, even if applying it to generate solutions to moral problems is often not a practical possibility.

The Divisiveness Objection

Another objection is that theological voluntarism is bound to be a divisive point of view. William K. Frankena puts the point this way:

> However deep and sincere one's own religious beliefs may be, if one reviews the religious scene, contemporary and historical, one cannot help but wonder if there is any rational and objective method of establishing any religious belief against the proponents of other religions or of irreligion. But then one is impelled to wonder also if there is anything to be gained by insisting that all ethical principles are or must be logically grounded on religious beliefs. For to insist on this is to introduce into the foundation of any morality whatsoever all of the difficulties involved in the adjudication of religious controversies, and to do so is hardly to encourage hope that mankind can reach, by peaceful and rational means, some desirable kind of agreement on moral and political principles
>
> (Frankena 1973: 313).

Though Frankena is in this passage discussing views in which the relation between religion and morality is logical, presumably he would have a similar worry about our theory in which the relation is metaphysical. And, of course, Frankena is correct in pointing out that religious disagreement has in the past given rise to moral disagreement and continues to do so.

But religious disagreement does not inevitably give rise to disagreement about moral principles. A theological voluntarist can agree with a secular Kantian deontologist on the principle that torture of the innocent is always morally wrong. They will, to be sure, disagree about why torture of the innocent is always wrong. A theological voluntarist who adopts our theory will say that it is wrong because God antecedently intends that no one ever bring about the torture of an innocent person. A secular Kantian deontologist may say that it is wrong because it involves failing to treat the humanity in another as an end in itself. Disagreement at the level of the metaphysics of morals is consistent with overlapping consensus at the level of moral principles. So despite religious disagreement, there are grounds for hope that we can reach, by peaceful and rational means, agreement on at least some moral and political principles.

It would, I think, be unrealistic to expect overlapping consensus on all matters of moral and political principle as long as disagreement in moral theory persists. But as Adams (1993) points out, nothing in the history of modern secular moral theory gives us reason to expect that general agreement on a single comprehensive moral theory will ever be achieved or that, if achieved, it would long endure in a climate of free inquiry. His conclusion, with which I agree, is that "the development and advocacy of a religious ethical theory, therefore, does not destroy a realistic possibility of agreement that would otherwise exist" (Adams 1993: 93). Philosophers should respond to disagreement in the area of moral theory in the same way they respond to theoretical disagreement in other areas, namely, as an opportunity to search for reasons that will reduce it. Their goal at every stage of the search should be agreement only to the extent that it is supported by the best available reasons. The mere fact that a moral theory provokes disagreement gives us no reason not to accept or advocate it. Nor does the fact that such theoretical disagreement may be an obstacle to reaching agreement on how to solve practical problems give us such a reason. So even if theological voluntarism is, to some

extent, a divisive point of view, this does not show that it is false or unworthy of serious consideration by moral theorists. And, to be fair, I should note that Frankena would probably agree with this conclusion; he acknowledges that if the view that morality is dependent on religion rests on good grounds, we must accept it (Frankena 1973: 314).

It is also worth noting that not all moral disagreement is divisive. A Kierkegaardian Christian may think that Mother Teresa was only doing her duty toward her neighbor as specified by the Love Commandment and regret that he fails to live up to the standards she set. One of her secular admirers may believe that much of the good she did was supererogatory. But if they agree that she did a great deal of good and that the world would be a better place if it contained more people like her, their disagreement about whether some good things she did were obligatory or supererogatory is not apt to be especially divisive.

The Anything Goes Objection

Perhaps the most troublesome objection to theological voluntarism was clearly stated by Ralph Cudworth. He said that

> divers Modern Theologers do not only seriously, but zealously contend . . ., *That there is nothing Absolutely, Intrinsically, and Naturally Good and Evil, Just and Unjust, antecedently to any positive Command of God; but that the Arbitrary Will and Pleasure of God*, (that is, an Omnipotent Being devoid of all Essential and Natural Justice) *by its Commands and Prohibitions, is the first and only Rule and Measure thereof.* Whence it follows unavoidably that nothing can be imagined so grossly wicked, or so foully unjust or dishonest, but if it were supposed to be commanded by this Omnipotent Deity, must needs upon that Hypothesis forthwith become Holy, Just and Righteous.
>
> (Cudworth 1731/1976: 9–10)

Consider some foully unjust state of affairs, say, an innocent child's being tortured to death. Translated into the idiom of our theory, Cudworth's complaint would be that theological voluntarism has as a consequence the following conditional:

(11) If God were antecedently to intend that someone at some time bring about the torture to death of an innocent child, then it would be morally obligatory for that person at that time to bring about the torture to death of an innocent child.

Cudworth is right about this point. Our theory's principle of obligation, (P1), has (11) among its consequences. But this will yield a successful refutation of our theory only if it can be shown that (11) is false. In order to show that (11) is false, one must show that its antecedent is true and its consequent is false. Can this be done?

There is a very plausible claim that entails the falsity of the consequent of (11). It is this:

(12) There is no possible world in which it is morally obligatory for anyone at any time to bring about the torture to death of an innocent child.

And the following claim entails the truth of the antecedent of (11):

(13) There is a possible world in which God antecedently intends that someone at some time bring about the torture to death of an innocent child.

But a theological voluntarist who accepts (12) can reject (13). A theological voluntarist can consistently reject the claim Cudworth makes parenthetically that God is an omnipotent being devoid of all essential and natural justice. If God is essentially just, there will be constraints on the antecedent intentions God can form. If it is unjust to bring about a certain state of affairs, it is also unjust to intend that anyone else bring it about. Hence, a theological voluntarist can maintain that there is no possible world in which God antecedently intends that someone at some time bring about the torture to death of an innocent child.

Theological voluntarists who are convinced that God is essentially just thus have a straightforward response to the objection. It is to admit that (11) is a consequence of their view but to insist that its antecedent is impossible. According to most theories of counterfactual conditionals, counterfactuals with impossible antecedents are trivially true. Thus theological voluntarists can accept (11) and hold that it is true. So the objection fails to refute theological voluntarism. In morality, it is not the case that anything goes if morality depends on the will of an essentially just God.

It might seem that it is not legitimate for theological voluntarists to appeal to divine justice. This would be the case if the only way to understand divine justice were in terms of obedience to certain self-addressed divine commands or fulfillment of certain intentions for divine action, for such things provide no constraints on the antecedent intentions God can form. But an alternative view is open to the theological voluntarist. It is that, while in the human case justice is both good and made obligatory by God, in the divine case justice is good but not obligatory. God's essential perfect goodness entails God's essential justice. So though God is not under an obligation to be just, God is just by a necessity of the divine nature. It is the divine nature itself, and not divine commands or intentions, that constrains the antecedent intentions God can form.

Of course, a theological voluntarist can also consistently accept (13) and reject (12). The discussions of the immoralities of the patriarchs by Augustine, Andrew, and Aquinas provide a precedent for this move. A theological voluntarist who takes this tack can accept (11) and hold that it is true because both its antecedent and its consequent are true at the appropriate possible world or worlds. In my opinion, this response to the present objection is less plausible than the response previously

considered. However, I think it would be a mistake to generalize to the conclusion that it is an implausible kind of response in every possible case, including all the cases of the immoralities of the patriarchs. Hence I do not think the contribution those examples make to my cumulative case for theological voluntarism is undercut by my preference for the first response to Cudworth's objection.

My strategy in responding to objections has been to rebut them one at a time. This seems to me fair because they are presented in this fashion by authors who criticize theological voluntarism. But, of course, someone might try to build a cumulative case against theological voluntarism by combining several objections. For example, I think Cudworth's objection would show promise of contributing to such a cumulative case if the second response to it I have discussed were the only response available to the theological voluntarist. However, I do not think the other objections I have considered show similar promise. So while I acknowledge that it is incumbent on defenders of theological voluntarism to give a hearing to and to try to rebut a cumulative case argument against their position if one is presented, I do not think that there is at present such a case to answer.

In sum, theological voluntarism is a view of the deontological part of morality that can be formulated with precision, supported from within a monotheistic worldview by a strong cumulative case argument, and defended against numerous objections. Thus our theory should be very attractive to ethical theorists who are monotheists. It should also command respect from ethical theorists who, while not themselves monotheists, are not hostile to monotheism.[1]

Note

1 I am grateful to Hugh LaFollette for helpful comments.

References

Adams, R.M. (1973) "A Modified Divine Command Theory of Ethical Wrongness," in *Religion and Morality*, eds. G. Outka and J.P. Reeder Jr, Garden City, NY: Anchor, pp. 318–47.

Adams, R.M. (1979) "Divine Command Metaethics Modified Again," *Journal of Religious Ethics* 7: 66–79.

Adams, R.M. (1993) "Religious Ethics in a Pluralistic Society," in *Prospects for a Common Morality*, eds. G. Outka and J.P. Reeder Jr, Princeton, NJ: Princeton University Press, pp. 93–113.

Adams, R.M. (1996) "The Concept of a Divine Command," in *Religion and Morality*, ed. D.Z. Phillips, London: Macmillan, pp. 59–80.

Andrew of Neufchateau (1514/1997) *Questions on an Ethics of Divine*, trans. J.M. Idziak, Notre Dame, IN: University of Notre Dame Press.

Aquinas, T. (1273/1948) *Summa Theologica*, trans. Fathers of the English Dominican Province, New York: Benziger.

Augustine of Hippo (426/1958) *The City of God*, trans. G.G. Walsh, D.B. Zema, G. Monahan, and D.J. Honan, Garden City, NY: Image.

Bentham, J. (1789/1948) *An Introduction to the Principles of Morals and Legislation*, New York: Hafner.

Cudworth, R. (1731/1976) *A Treatise Concerning Eternal and Immutable Morality*, New York: Garland.

D'Arcy, E. (1973) "'Worthy of Worship': A Catholic Contribution," in *Religion and Morality*, eds. G. Outka and J.P. Reeder Jr, Garden City, NY: Anchor, pp. 173–203.

Frankena, W.K. (1973) "Is Morality Logically Dependent on Religion?" in *Religion and Morality*, eds. G. Outka and J.P. Reeder Jr, Garden City, NY: Anchor, pp. 295–317.

Idziak, J.M. (1989) "In Search of 'Good Positive Reasons' for an Ethics of Divine Commands: A Catalogue of Arguments," *Faith and Philosophy* 6: 47–64.

Idziak, J.M. (1997) "Divine Command Ethics," in *A Companion to Philosophy of Religion*, eds. P.L. Quinn and C. Taliaferro, Oxford: Blackwell, pp. 453–9.

Kierkegaard, S. (1847/1964) *Works of Love*, trans. H.V. Hong and E.H. Hong, New York: Harper.

Loux, M.J. (1986) "Toward an Aristotelian Theory of Abstract Objects," *Midwest Studies in Philosophy* 11: 495–512.

Morris, T.V. (1987) *Anselmian Explorations*, Notre Dame, IN: University of Notre Dame Press.

Mouw, R.J. (1990) *The God Who Commands*, Notre Dame, IN: University of Notre Dame Press.

Murphy, M. (1998) "Divine Command, Divine Will, and Moral Obligation," *Faith and Philosophy* 15: 3–27.

Quinn, P.L. (1978) *Divine Commands and Moral Requirements*, Oxford: Clarendon Press.

Quinn, P.L. (1979) "Divine Command Ethics: A Causal Theory," in *Divine Command Morality: Historical and Contemporary Readings*, ed. J.M. Idziak, New York and Toronto: Edwin Mellen Press, pp. 305–25.

Quinn, P.L. (1990a) "An Argument for Divine Command Ethics," in *Christian Theism and the Problems of Philosophy*, ed. M.D. Beaty, Notre Dame, IN: University of Notre Dame Press, pp. 289–302.

Quinn, P.L. (1990b) "The Recent Revival of Divine Command Ethics," *Philosophy and Phenomenological Research* 50: 345–65.

Quinn, P.L. (1992) "The Primacy of God's Will in Christian Ethics," in *Philosophical Perspectives 6*, ed. J.E. Tomberlin, Atascadero, CA: Ridgeview, pp. 493–513.

Quinn, P.L. (1996) "The Divine Command Ethics in Kierkegaard's *Works of Love*," in *Faith, Freedom, and Rationality*, eds. J. Jordan and D. Howard-Snyder, Lanham, MD: Rowman & Littlefield, pp. 29–44.

Wierenga, E.R. (1989) *The Nature of God: An Inquiry into Divine Attributes*, Ithaca, NY: Cornell University Press.

Chapter 5

Moral Intuition

Jeff McMahan

Moral Inquiry

Suppose we wish to understand a particular moral problem – for example, abortion. One way of proceeding is to reason on the basis of our existing substantive moral beliefs. We may, however, suspect that our moral beliefs about abortion, insofar as we have any prior to serious reflection, are unreliable. We may suspect, for example, that our beliefs about abortion reflect the influence of a religious education that we now repudiate, or that our feminist sympathies may make us insufficiently sensitive to the status of the fetus. Thus we may take as our starting point certain related moral beliefs about which we are more confident – for example, that killing an unthreatening, morally innocent, adult human being (or "person," for the sake of brevity) whose continued life would be worth living is, except perhaps in the most extreme circumstances, seriously wrong, while painlessly killing a lower nonhuman animal (for example, a mouse) may be permissible provided that the interests that are thereby served outweigh those of the animal that would be frustrated by its death. There is, of course, divergence of opinion even about these cases. Some people believe that intentionally killing a person can never be justified, while others believe that it can be justified whenever it is necessary to save the lives of a greater number of innocent people. And, while some believe that there is no objection whatever to killing a mouse independently of the effects this might have on human interests, others believe that killing a mouse is seriously objectionable just because of the effect on the mouse and requires a strong justification in order to be permissible. Nevertheless, everyone agrees that, in ordinary circumstances, killing a person is more objectionable morally than killing a mouse.

The Blackwell Guide to Ethical Theory, Second Edition. Edited by Hugh LaFollette and Ingmar Persson.

We could therefore initiate our inquiry into abortion by exploring our confident sense that there is a significant moral difference between killing a person and killing a lower animal – so that, for example, the killing of a lower animal might be justified by appeal to considerations that would not constitute even the beginning of a justification for killing a person. With these beliefs as our starting point, we could work our way toward a better understanding of abortion. We could proceed by trying to understand why killing a person is generally wrong and why it is generally so much more seriously wrong than killing a lower animal. What are the relevant differences between a person and a lower animal? Are the properties of persons that make killing them generally worse all intrinsic properties? Or is part of the explanation of the greater wrongness of killing persons that we normally bear certain relations to them that do not exist between ourselves and animals? In addressing these questions, we may consult our intuitions about a range of particular cases and this may yield provocative results. We may notice, for example, that the extent to which killing an animal seems wrong varies with the degree of harm the animal suffers in dying. Thus it seems more objectionable to kill a dog than to kill a mouse; and the explanation seems connected with the fact that the dog loses more by dying. But we may also notice that the extent to which it is wrong to kill a person does not seem to vary with the extent to which death is bad for the victim. Thus it seems no less wrong, other things being equal, to kill a dullard than to kill a genius, or to kill an elderly person with a reduced life expectancy than to kill a person in the prime of life.

As our understanding of the morality of killing in general increases, we can begin to extract from our findings various implications for the morality of abortion. Suppose, for example, that we discover that there are certain properties that adult human beings generally possess that lower animals do not that seem to help explain the difference between killing people and killing animals. We can then consider whether these properties are possessed by human fetuses. If they are, then in that respect abortion is relevantly like the killing of a person; if not, there is then reason to suppose that abortion is morally more like the killing of animals.

These remarks about abortion are intended only to provide a sketchy illustration of a certain approach to practical ethics, a certain general pattern of reasoning about moral problems. Its most conspicuous feature is that it treats certain substantive moral beliefs that we already have as prima facie reliable starting points for moral inquiry. It presupposes that at least some of our moral intuitions have a certain presumptive epistemic authority.

Intuitions

What are moral intuitions? As I will understand the term, a moral intuition is a moral judgment – typically about a particular problem, a particular act, or a particular agent, though possibly also about a moral rule or principle – that is not the

result of inferential reasoning. It is not inferred from one's other beliefs but arises on its own. If I consider the act of torturing the cat, I judge immediately that, in the circumstances, this would be wrong. I do not need to consult my other beliefs in order to arrive at this judgment. This is not to say that a moral intuition is necessarily elicited instantaneously, the way a sense perception is. If a particular problem or case is complex, one may have to consider it at length in order to distinguish and assimilate its various relevant features – in much the same way that one might have to examine the many details of a highly complex work of art in order to judge or appreciate it.

The belief that I cited as one of the possible starting points for an inquiry about abortion – namely, that killing a person is generally wrong – may or may not count as an intuition according to this understanding, though for most of us it has an intuitive basis. It is an intuition if one finds it immediately compelling, but not if one accepts it as an inductive inference from one's intuitively finding that in this, that, and the other case, killing a person is wrong, or if one deduces it from the principle that whatever God prohibits is wrong.

In the history of moral philosophy, the idea that moral intuitions have epistemic authority has been associated, unsurprisingly, with a cluster of theories that have traveled under the label "intuitionism." Those doctrines are many and various, and I do not propose to disentangle them. But two claims associated with certain historically prominent variants of intuitionism have done much to discredit the appeal to intuitions. One is that intuitions are the deliverances of a special organ or faculty of moral perception, typically understood as something like an inner eye that provides occult access to a noumenal realm of objective values. The other is often regarded as a corollary of the first – namely, that intuitions are indubitable (that is, that their veridicality cannot be doubted) as well as infallible (that is, that they cannot in fact be mistaken). But a variety of considerations – such as the diversity of moral intuitions, the fact that people do often doubt and even repudiate certain of their intuitions, and the evident origin of some intuitions in social prejudice or self-interest – make it untenable to suppose that intuitions are infallible apprehensions of moral reality by some special faculty of moral perception.

There are other features that are occasionally attributed to intuitions that are in fact inessential. It is sometimes said, for example, that intuitions are "pretheoretical." If all this means is that they are not derived inferentially from a moral theory, then this of course follows from the stipulation that they are not the products of any sort of inferential reasoning. If, however, it means that intuitions must be untutored or entirely unaffected by a person's exposure to moral theory, then the requirement is evidently too strong. Just as many people's moral intuitions have been shaped by their early exposure to religious indoctrination, so some people's intuitions are gradually molded by their commitment to a particular moral theory. The stipulation that intuitions are not the products of conscious inference does not entail that they cannot be affected by learning. That they may arise spontaneously is compatible with their having sources in one's nature that are malleable.

Theory

Many philosophers reject the idea that moral intuitions have epistemic authority. Peter Singer, for example, has suggested that we should assume "that all the particular moral judgments we intuitively make are likely to derive from discarded religious systems, from warped views of sex and bodily functions, or from customs necessary for the survival of the group in social and economic circumstances that now lie in the distant past." On this assumption, he notes, "it would be best to forget all about our particular moral judgments" (Singer 1974: 516). It is, of course, possible to be less dismissive of intuitions and yet still regard them as lacking in authority. Some philosophers, for example, concede that intuitions may be reasonably reliable guides to action in most circumstances – since morality must ensure that people are equipped with dispositions to believe and act in certain ways in situations in which deliberation and reflection are not possible – but deny that they are a source of moral knowledge or have any proper role in reasoning about moral problems. They believe that practical reasoning about a moral problem must consist in determining what some favored moral theory implies about the problem. It is the theory that is the source of our moral knowledge concerning particular problems and cases. And the theory is itself validated by means other than its conformity with intuition.

According to this approach, if our concern is to understand the morality of abortion, our first task must be to discover the correct moral theory. Moral inquiry is initially and primarily theoretical; only at the end of this theoretical inquiry is it possible to address moral problems such as abortion competently, bringing the theory to bear and extracting from it the knowledge we initially sought. This general approach therefore contrasts with the first approach I sketched, according to which moral inquiry begins with problems and cases and our intuitions about them, seeks principles that unify and explain the intuitions, and proceeds through adjustment and modification of both the principles and intuitions until consistency and harmony are achieved. On this approach, a moral theory in which we are entitled to have confidence is something that we can hope to have only near the end of the process of inquiry into problems of substantive morality.

Let us refer to the two broadly defined patterns of moral inquiry that I have sketched as the *Intuitive Approach* and the *Theoretical Approach.* Both are richly represented in the history of moral philosophy. The Socrates of Plato's dialogues is an admirable exemplar of the Intuitive Approach, while Hobbes and Kant exemplify the Theoretical Approach. Each of the latter begins with a conception of the nature of morality that he believes dictates a particular method for arriving at moral judgments about particular problems and cases. In recent years, most philosophers working on problems of practical ethics have largely followed the Intuitive Approach, but the Theoretical Approach also has many distinguished recent exponents, among them Richard Hare, Richard Brandt, and an assortment of theorists in the contractualist and consequentialist traditions.

The Theoretical Approach is reformist in a rather radical way. People have always reasoned and argued about substantive issues in morality. According to adherents of the Theoretical Approach, however, people have been misguided insofar as their reasoning has diverged from the forms and patterns of moral reasoning prescribed by the correct moral theory. Richard Brandt, for example, suggests that "is morally wrong" means "would be prohibited by any moral code which all fully rational persons would tend to support, in preference to all others or none at all, for the society of the agent, if they expected to spend a lifetime in that society" (Brandt 1979: 194). Assuming that this definition also states a test for determining whether an act is wrong, it seems clear that any convergence of the conclusions of most people's actual moral reasoning and the conclusions that might result from Brandt's proposed mode of reasoning would be to a considerable degree fortuitous or coincidental. According to the Theoretical Approach, therefore, philosophical ethics is utterly different from, say, the philosophy of science. While the philosopher of science may criticize certain aspects of the practice of science, and may urge scientists to revise their understanding of the nature of their practices or the status of their conclusions, the philosopher does not presume to tell scientists that they have been utilizing the wrong method and would do better to adopt a different approach. The Intuitive Approach is in general more respectful of the modes of moral reasoning that people actually employ – though only because people in fact tend to reason about moral problems in the way it recommends.

Theory Unchecked by Intuition

Could we really conduct our thinking about morality and moral problems in the way suggested by the Theoretical Approach, without building up from our moral intuitions or consulting those intuitions to test the plausibility of the implications of proposed moral theories? Even those who most vehemently deny that intuitions have any independent credibility nevertheless often build their arguments on the basis of appeals to common intuitions (for example, Rachels 1986: 112–13, 134–5; Singer 1993: 229). But, although this is suggestive of the difficulty of producing arguments capable of persuading people without linking them to our preexisting moral beliefs, it is merely an ad hominem point and as such does little to support the appeal to intuition. An alternative point that may be urged against the Theoretical Approach is that our intuitions often *compel* belief in a way that, for most of us, no moral theory does. If an intuition that is highly compelling cannot be reconciled with what seems to be the best supported moral theory, can it be rational to reject the intuition at the behest of a theory that is feebler in its ability to compel belief?

It is important to be clear about the nature of this challenge. The claim is not simply that moral intuitions often strike us as more obvious or less open to

doubt than it seems any moral theory is. By itself, this would not be a strong consideration in favor of the intuition. The theories of modern physics tell us that many of our commonsense beliefs about the nature of the physical world are mistaken. Many of these beliefs seem overwhelmingly obvious while the theory that disputes them may be so arcane as to be unintelligible to all but a few. Yet most of us recognize that some scientific theories that overturn aspects of our commonsense conception of the physical world are so well established by their powers of explanation and prediction and by the control they give us over the forces of nature that we readily acquiesce in their claims and concede that our commonsense views must be illusory. If a moral theory could command our allegiance by comparable means of persuasion, we might yield our intuitions to it without demur, even if it had none of the immediate obviousness in which our intuitions tend to come clothed. But the challenge to the Theoretical Approach is that no moral theory, at least at the present stage of the history of philosophical ethics, can have the degree of authority or validation that the best supported scientific theories have. The lamentable truth is that we are at present deeply uncertain even about what types of consideration support or justify a moral theory. There are no agreed criteria for determining whether or to what extent a moral theory is justified. So when an intuition, which may be immediately compelling, comes into conflict with a moral theory, which can have nothing approaching the authority of a well-grounded scientific theory, it is not surprising that we should often be reluctant to abandon the intuition at the bidding of the theory. We can, indeed, be reasonably confident *in advance* that none of the moral theories presently on offer is sufficiently credentialed to make it rationally required that we surrender our intuition.

It is instructive to consider how most of us respond when, on inquiring into a particular moral problem, we find that some moral theory has implications for the problem that clash with our intuitions. Our response is not to question how well grounded the theory is, on the assumption that we should be prepared to acquiesce if we find that the theory is well supported. If the theory generates its conclusion via a distinct argument, our tendency is to detach the argument from the parent theory and consider it on its own merits. According to R.M. Hare, for example, his universal prescriptivist theory of morality implies that we should reason about the morality of abortion by applying a variant of the Golden Rule: "We should do to others as we are glad that they did do to us" (Hare 1975: 208). When we discover that, at least according to Hare, this principle implies not only that abortion is generally wrong but also that remaining childless is as well, we do not go back to Hare's earlier books to check the arguments for universal prescriptivism. Instead we undertake an independent inquiry to try to determine whether and, if so, to what extent it matters to the morality of abortion that, when an abortion is not performed, there will typically later be a person who is glad to exist who would not have existed if the abortion had been performed. That is, if we are serious about understanding the morality of abortion, we will take seriously the considerations identified as relevant by the theory; and we may be grateful to

the theory for having helped us to see whatever relevance these considerations may in fact have; but we are generally not overawed by the fact that these considerations have been identified as relevant *by the theory*. Their provenance in the theory fails to imbue them with epistemic authority.

One might even wonder what claim a theory might have to be a *moral* theory if it has foundations that are wholly independent of the intuitions that have shaped the common features of all recognizably moral codes. Recall, for example, Brandt's claim that an act is morally wrong if it "would be prohibited by any moral code which all fully rational persons would tend to support, in preference to all others or none at all, for the society of the agent, if they expected to spend a lifetime in that society." It is possible that fully rational persons would tend to prefer the code that, if generally accepted, would make them (though not necessarily nonrational beings, such as animals) most comfortable. The methodology rules out the possibility that, in choosing such a code, they would be failing to recognize or respect their own moral status. (Recognition of their moral status could not, of course, be a condition of their being rational, for that would make the account circular.) I cite Brandt's theory for the purpose of illustration because it imposes fewer constraints on the choice of principles than most other contractualist and rule-consequentialist theories, and is therefore less likely than they are to yield principles that support common intuitions. But a similar concern arises for all theories that reject conformity with intuition as a constraint on moral justification.

Moral Epistemology

The remarks in the previous section are meant only to suggest certain reservations we might have about the Theoretical Approach; they are far from providing decisive reasons for rejecting that approach. Moreover, even if we had stronger grounds for skepticism about the Theoretical Approach than those I have offered, this would still be insufficient to compel us to accept the Intuitive Approach. For it is hardly a ground for confidence in our intuitions that there are reasons for doubting the approach to moral inquiry that denies them a role. Something more positive has to be said on behalf of our intuitions themselves. At a minimum, more needs to be said about the role they are supposed to have in the structure of justification in ethics. In what follows I will first offer a few general remarks about moral epistemology, after which I will briefly sketch an account of moral inquiry that explains the role that our intuitions might have in our moral thinking and also helps to elucidate their epistemological status. I will then conclude by noting that there is a conception of the nature of moral knowledge that has independent plausibility and, if correct, offers a deeper understanding of the epistemological status of moral intuitions.

A theory in epistemology may be a theory either of truth or of justification. I will focus on the issue of justification and simply assume that there is a tight

connection between justification and truth. Accounts of justification, including accounts of justified moral belief, tend to be divided into two major approaches: coherentism and foundationalism. Coherentist accounts of moral justification hold that a moral belief is justified solely in terms of its relations, particularly its inferential relations, with other beliefs. It is justified to the extent to which it coheres with a set of beliefs that together form a coherent whole. By contrast, foundationalist accounts hold that some beliefs are self-justifying – at least in the sense that they are justified independently of their relations to other beliefs. According to foundationalist accounts, a moral belief is justified if and only if it is either self-justifying or bears an appropriate inferential relation to a belief that is self-justifying.

Of the two types of account, coherentism is generally thought to be more hospitable to the Intuitive Approach. The most commonly endorsed method of moral inquiry among contemporary moral philosophers is the method described by John Rawls under the label "reflective equilibrium" (Rawls 1972: 19–21, 48–51). According to the method of reflective equilibrium, we begin with a set of moral intuitions about particular cases, filter out those that are the obvious products of distorting influences, and then seek to unify the remaining intuitions under a set of more general principles. We seek principles that both imply and explain our particular judgments. But the match between principles and intuitions will inevitably be very imperfect in the first instance. A candidate principle may imply a great many of our intuitions and yet have some implications that conflict with other intuitions. In that case we may modify or even abandon the principle; but, if the principle has considerable explanatory power with respect to a wide range of intuitions and cannot be modified without significant sacrifice of this power, we may instead decide to reject the recalcitrant intuitions. In this way we make reciprocal adjustments between intuitions and principles until our beliefs at various levels of generality are all brought into a state of harmony, or reflective equilibrium. This method is generally interpreted in coherentist terms, in that it is understood to make coherence with other beliefs the sole criterion of a belief's credibility. Yet it obviously treats intuitions as potential sources of moral knowledge. Although intuitions arise noninferentially and thus, in coherentist terms, have no prima facie credibility on their own, those that survive the initial filtration and are compatible with the principles that emerge in the process of trying to reach reflective equilibrium turn out to be justified moral beliefs.

Foundationalist theories of moral justification tend to be favored by proponents of the Theoretical Approach. Typically, the foundational beliefs (that is, those that are not justified in terms of their relations to other beliefs) are held to be nonmoral; justified moral beliefs are all ultimately derivable via some process of reasoning that is based on the foundational nonmoral beliefs (for examples, see Timmons 1987). Those who attribute authority to our moral intuitions tend, understandably, to be more reluctant to embrace foundationalism. This is mainly because it seems implausible to regard our intuitions themselves as foundational. This seems to attribute to them too exalted a status. While our intuitions do seem to have a

certain initial credibility, it seems exorbitant to suppose that they are self-evident or self-justifying. They seem at most to have an evidentiary status. We recoil from the suggestion (advanced, as I noted earlier, by various traditional intuitionists) that intuitions are the unshakable basis on which all moral knowledge rests.

There are, however, at least two ways of overcoming this ground of reluctance to combine foundationalism with the Intuitive Approach. The first is to recognize that a belief may be of the foundational *sort* and yet be defeasible. Suppose, for example, that sense perceptions are the foundations of empirical knowledge. Even if all empirical knowledge is derived immediately from sense perceptions or is ultimately traceable by chains of inference to sense perceptions, it does not follow that *all* sense perceptions are sources of empirical knowledge. Some may be distorted, illusory, or otherwise erroneous. And there is no reason why the same may not be true, *mutatis mutandis*, in the case of moral intuitions.

Second, a foundationalist account of moral knowledge may treat intuitions as reliable sources of moral knowledge without treating them as foundational or self-justifying. It is this possibility that I will explore in more detail.

A Sketch of a Foundationalist Conception of Moral Justification

To most moral philosophers who reason about substantive moral issues, it seems that the method of reflective equilibrium, or a process very similar to it, is the best or most fruitful method of moral inquiry. Of the known methods of inquiry, it is the one that seems most likely to lead to justified moral beliefs. It does not, however, have to be interpreted within the coherentist framework. It is compatible with a foundationalist conception of moral justification.

Here, in outline, is a more detailed account of how the method works. Again, we begin with intuitions about particular problems, particular cases. If our initial interest is in a problem about which we have no intuitions, or about which our intuitions are weak or conflicting, we should, as I have suggested with reference to the problem of abortion, find closely related cases about which we have confident intuitions and work from these. The question arises, however, why we should carry the inquiry any further. Why cannot we rest content with our intuitions, allowing ourselves to be guided by them on a case-by-case basis? Part of the answer, of course, is that there are many moral problems about which we have no intuitions, or about which the intuitions we have are weak, conflicting, or obviously suspect or dubious. We need a method for determining what we should believe and what we should do in cases such as these.

When one's moral intuition is challenged by another person, it is natural to respond by appealing to claims of a higher level of generality that imply or explain the intuition. The assumption is that the credibility of the intuition is enhanced if it can be subsumed under a plausible moral principle. So, for example, the intuitive judgment that it would be wrong to torture the cat for fun might be defended by

appealing to the principle that it is wrong to cause suffering without a justifying reason. Private moral reflection may follow the same dialectical pattern as moral disagreement between persons. We should challenge our own intuitions in much the way that an opponent might challenge them; but we may also respond in much the same way, by trying to bring them within the scope of a plausible principle.

But why suppose that the credibility of one's intuition is enhanced when it is shown that the intuition is implied by the conjunction of a moral principle and the facts of the case? One suggestion is that the principle may elucidate the intuition by identifying the features of the case that are morally salient. If, for example, one feels intuitively that it is wrong to kill animals for sport, one's objection is sharpened or focused if it is seen to follow from the more general view that it is morally objectionable, in the absence of sufficient justification, to deprive any individual of a good that that individual would otherwise have. The principle brings out more clearly at least part of what one finds intuitively objectionable.

No one supposes, of course, that just any principle will do. For the principle to support the intuition, it must have independent credibility. The principle itself may have intuitive appeal. According to coherentism, the *mere* fact that the principle implies the intuition provides some minimal epistemic support for each; for mutual coherence among beliefs is the criterion of justifiability. But, even according to coherentism, the principle will provide no more than token support for the intuition unless it is itself well integrated within a larger network of mutually coherent beliefs. Hence the method of reflective equilibrium demands that the principle itself be tested for consistency and coherence with other intuitions and principles. Its implications about particular cases should not conflict with one's intuitive judgments about those cases and, to the greatest extent possible, its implications should not contradict the implications of other principles one accepts. It is, of course, too much to require that moral principles not have conflicting implications: conflict is the price of pluralism. But conflicts should, in principle, be resoluble, in that one recognizes the necessity of one value's yielding to another, though not without some irreducible loss.

So the defense of one's initial intuition by subsuming it under a more general principle is only the beginning of moral inquiry. The principle must itself be assessed by testing its implications for consistency with one's other beliefs. One need not accept coherentism in order to appreciate the importance of this test. There are practical reasons why inconsistent moral beliefs are problematic: they may, for example, provoke indecision and, ultimately, paralysis of the will. More importantly, the achievement of greater coherence among one's beliefs diminishes the likelihood of error by helping one to identify and eliminate moral beliefs produced by self-interest, faulty reasoning, failure of imagination, illusory metaphysical beliefs, impaired faculties, and other sources of distorted or mistaken belief.

But there is a deeper basis for trying to subsume an intuition under a principle that is itself supported by its power to unify and explain a range of other intuitions. This is that the process of achieving increasing coherence among principles and intuitions facilitates the discovery of deeper values and also brings surface beliefs

about particular cases into alignment with those deeper values in a way that reveals and illuminates the connections between them. When one seeks to formulate a moral principle that implies and illuminates a widespread and robust intuition about a particular problem or case, one is in fact groping or probing for deeper values of which it is a surface manifestation. The expectation that the principle will illuminate and explain the intuition assumes that the intuition is in fact an expression, in a particular context, of a value that is deeper, more basic, and more general than the intuition itself. One's efforts to formulate the principle and to revise and refine it in a way that brings more and more common intuitions within its scope are attempts to capture or articulate some core principle in its full generality, to get its form exactly right, omitting nothing, however subtle.

This process, as I have described it, is indistinguishable from that endorsed by the coherentist practitioners of reflective equilibrium. One seeks support for an intuition by appealing to a principle, then one seeks to support the principle by demonstrating its compatibility with other intuitions, and so on. Why not understand the method as most people do, in coherentist terms?

There are various general objections to coherentist accounts of justification in ethics (e.g., Gaus 1996: ch. 6). I will not rehearse them here. (There are also, of course, general objections to foundationalism; I will not discuss those either.) I will simply note two problems that I find particularly disturbing. One is that, according to coherentism, no belief is immune to rejection, no matter how compelling it may be. If its elimination from the network of beliefs would enhance overall coherence within the network, the belief must go. Indeed, it seems possible, though not likely, that a coherentist approach to the pursuit of reflective equilibrium could lead ultimately to the rejection of every belief with which one started. Both these suppositions, however, are alien to moral life and moral reflection. There are some moral beliefs that we simply cannot give up just for the sake of greater coherence. Sometimes we must hold tenaciously to certain convictions even when it seems that greater coherence or systematicity could be achieved by rejecting them. According to the coherentist, one's moral beliefs are like pieces in a game: one shuffles them around, sacrifices some, and acquires others, all for the sake of achieving certain relations among them. No piece has any significance in itself; it has significance only in relation to the other pieces and in particular in the contribution that it makes to the whole of which it is a part. If moral reflection were really a game like this, in which our moral beliefs had no claim to commitment from us and thus were always expendable in the service of coherence, coherence would be easily achievable. It is because some of our moral beliefs compel our allegiance independently of their inferential relations to other beliefs that full coherence always seems such a distant goal.

A closely related worry about coherentism is that it assigns the same epistemic status to our intuitions about particular cases that it assigns to the deeper principles of which the intuitions are expressions. They stand in relations of reciprocal support: the principles imply the intuitions and we can therefore infer our way to the principles on the basis of the intuitions. But in fact the relations of reciprocal

support seem asymmetrical: the principles seem to be epistemically more basic, more secure. They articulate our core values which unify, explain, and justify our intuitive judgments. Our intuitions do not so much justify the principles as provide evidence of their existence and guidance as to their nature. In short, the principles are foundational with respect to the intuitions. Insofar as our intuitions are reliable sources of moral knowledge, they are so because they are expressions of, and point back to, a range of deeper, more general values.

As I noted earlier, foundationalism is distinguished by the view that some beliefs are justified independently of their relations to other beliefs. I will refer to these beliefs as "foundational." Among those who accept a foundationalist moral epistemology, there is a rough division between those who take certain nonmoral beliefs to be foundational and those who identify certain moral beliefs as foundational. Among the latter, there is a further division between those who take intuitions to be foundational (e.g., Ross 1930 and Gaus 1996) and those who take some general principle or principles to be foundational (e.g., Sidgwick 1907). In Geoffrey Sayre-McCord's words:

> Many have treated the privileged [that is, foundational] moral beliefs as roughly on a par with perceptual judgments and suggested that the justification of our various moral principles parallels the kind of justification our scientific principles receive from perception. Others have thought that our privileged moral beliefs concern, instead, the most general and abstract principles of morality, and that these in turn serve to justify (or not) our other beliefs deductively.
>
> (Sayre-McCord 1996: 150)

The view that I have suggested is of this second sort. But it is distinguished from many views of this sort in that it does not regard the foundational principles as self-evident or accessible directly through the exercise of intuition. Many philosophers (such as Sidgwick 1907 and Unger 1996) have regarded our intuitions about principles as more reliable than our intuitions about particular cases. But, insofar as a moral principle is substantive in character rather than merely formal (for example, "Treat like cases alike"), it seems a mistake to have confidence in our intuitive apprehension of the principle. To be justified in accepting a moral principle, we must first understand what it commits us to in particular cases. As William James noted in a letter written long before he became a practicing philosopher, "No one sees farther into a generalization than his own knowledge of the details extends" (Barzun 1983: 14). So, while I regard the principles rather than our intuitions as foundational, I do not think that moral inquiry can proceed by deducing conclusions about particular cases from self-evident moral principles. Rather, *the order of discovery is the reverse of the order of justification*. Although the deeper principles are explanatorily prior, we have to work our way to them via our intuitions in much the way that scientists work towards general principles via perceptual data. The process of discovering and formulating the more general principles is evidently difficult and intellectually demanding, rather in the way that

discovering the syntactic structures that govern our use of language is. As this familiar analogy suggests, as we grope our way towards the principles, we may be discovering what we antecedently believe, albeit below the level of conscious awareness, or discovering truths of which we were unaware. The principles that we hope to uncover may express deep dispositions of thought and feeling that operate below the level of consciousness to regulate our intuitive responses to particular cases.

If this is right, it explains why we experience the process of moral inquiry as a process of discovery rather than an exercise of choice or will. It also explains why the foundationalist approach I am describing should coincide with coherentism in holding that we may expect to arrive at a moral theory, if at all, only near the end of reflection about particular problems and cases rather than coming to the problems with a theory ready to be "applied." It explains why it is suspect when philosophers emerge from graduate school already believing themselves to be in possession of the correct moral theory.

Challenges

This brief sketch of an account of moral justification raises far more questions than can be answered, or even addressed, in the remainder of this short essay. The central question is, of course, what reason is there to suppose that moral intuitions, even those that are widespread and robust, are reliable guides to the formulation or discovery of principles that in turn provide reliable foundations for particular moral judgments? If the principles toward which common intuitions seem to direct us turned out to be luminously self-evident, this could reinforce our confidence in the intuitions. But that seems unlikely. Thus far the efforts of philosophers to find general principles that unify and provide a deeper foundation for common intuitions have led them to principles that have little immediate appeal on their own – for example, the principle that the intention with which one acts, or the causal relations among the consequences of one's act, matter to the permissibility of the act. Unlike the principle that one ought to do the act that would have the best consequences, impartially considered (which may seem plausible *until* one tests its implications against common intuitions), these principles seem attractive primarily because they promise to provide some underlying unity to our intuitions. They may not seem, therefore, to provide independent support for the intuitions, apart from showing how they cohere with one another.

There are, moreover, numerous objections to according any epistemic authority to moral intuitions, however common or robust. Perhaps the easiest of these objections to rebut is that there is an "inbuilt conservatism" (Singer 2005: 347) in any method of moral justification that gives an important role to moral intuitions. This concern goes back at least to Mill, who, in Alan Ryan's words, thought that reliance on intuitions in ethics "amounted to the sanctification of any opinion

that had been held long enough and deeply enough. It was the great intellectual buttress of social, moral, political, and intellectual conservatism" (Ryan 2011: 62). Yet the fact that each of us has many intuitions that, on inspection, conflict with others we hold with equal conviction, so that even common and robust intuitions frequently conflict with other equally common and robust intuitions, means that our efforts to achieve consistency among common intuitions are inevitably revisionist, sometimes radically so, when strong intuitions cannot be reconciled. Philosophers who accept that moral intuitions have a presumptive epistemic authority, and thus have employed the method I described in the earlier section "A Sketch of a Foundationalist Conception of Moral Justification," have consequently been driven by the demand for consistency to defend quite heterodox and indeed counterintuitive conclusions, such as that infanticide can be permissible in a variety of circumstances, that each person in wealthy societies ought to give the bulk of his or her wealth to people in impoverished societies, and so on.

Another familiar objection to the appeal to moral intuitions is that they are often tainted by their origin. Earlier I quoted Singer as observing that many common intuitions have their ultimate source in primitive religious beliefs, ancient taboos about sex, and social practices that were useful in a world that is no longer ours. There are many other tainted sources, such as superstitions concerning purity and defilement, and, perhaps most important, individual and collective self-interest. The vast majority of whites in the antebellum South thought it obvious that the enslavement of blacks was morally justified. Although they sought biblical and biological warrant for the practice, what really motivated their belief was crude self-interest. As long as they all had a strong interest in maintaining the practice and could reinforce each other's beliefs by participating in the practice and raising their children to accept it as part of the natural background to their lives, they were able to insulate their intuitive sense of the rectitude of the institution of slavery from challenges that would otherwise have disturbed it. People do the same today with their belief that it is permissible to kill animals in order to eat them. Because eating meat gives them pleasure, people assume that it is in their interest (though in the forms and quantities in which they consume it, it is not). Most people therefore eat meat and this itself shields them from critical reflection. For they assume that because virtually everyone does it, including the very nicest people they know, it simply cannot be seriously wrong to do it. Yet without the blinkering effects of self-interest, the many powerful moral objections to this practice would be obvious.

The fact that many common moral intuitions are discredited by their provenance is not, however, a problem if we have rejected the idea that intuitions result from the exercise of some special faculty for the direct inspection of moral reality. No one supposes that all moral intuitions, or even all "considered moral judgments," are correct. They are instead merely *appearances* (Huemer 2005: 99–105), some of which, we recognize, are bound to be delusions. According to the method of moral inquiry I described earlier, an essential part of the process of arriving at

foundational moral principles is the filtering out of intuitions that are contaminated at the source.

Some have argued, however, that the eradication of bias in our intuitions is impossible because recent work on the psychology of moral judgment and in "experimental philosophy" has revealed sources of systematic bias and irrationality in our moral intuitions that seem inherent in our psychological nature. This work has shown, for example, that our retributive intuitions become stronger and harsher in the presence of a bad smell, which is obviously irrelevant to whether or to what extent a person deserves to be harmed. Given that our intuitions are subject to this kind of distortion, it seems that they cannot be reliable guides to the discovery of foundational moral principles. For one aim of the method of moral justification I have described is to abstract from or transcend the idiosyncrasies of our personal psychologies or subjective points of view in order to achieve the greatest possible degree of objectivity and impartiality. Yet if common intuitions are systematically distorted by psychological mechanisms of which we are unaware, and that may not even be accessible to introspection, then we have presumably been failing to achieve objectivity even while doing all we can to succeed.

What this new work in psychology and experimental philosophy shows, however, is that we have not been doing all we could. For until just recently we have not been doing the invaluable work that psychologists and experimental philosophers have begun to do in identifying these subtle sources of distortion and irrationality in common moral intuitions. Even if these distorting influences cannot be eradicated in our intuitive thinking, they can, once they have been recognized, be controlled for in our moral theorizing. What experimental philosophers have forced us to see is that their empirical investigations are essential to the process of filtering raw intuitions and hence indispensable in the process of moral justification I have described.

A third concern about the epistemic status of moral intuitions arises, paradoxically, from their surprising uniformity over time and across cultures. We have been impressed for so long by the claims of cultural anthropologists, postmodern relativists, undergraduates, and others about the diversity of moral opinion that we have tended to overlook how much agreement there really is. Interestingly, moral disagreements seem to widen and intensify the more we abstract from particular cases and focus instead on matters of principle or theory. When the partisans of different schools of moral thought turn their attention to particular cases, there is far more intuitive agreement than their higher-level disputes would lead one to expect. As teachers of ethics will testify, there is often a remarkably high level of agreement among students about what it is permissible to do in specially contrived examples, such as the well-known trolley cases (which have figured heavily in much of the recent work on the psychology of moral judgment), even though the students may come from widely varying religious and cultural backgrounds and have never had any experience of problems of the sort about which they are invited to offer an intuitive judgment. What accounts for the agreement?

A potentially debunking explanation is that it has a biological basis. It may be, as Singer puts it, that "our common evolutionary heritage has, unsurprisingly, given us a common set of intuitive ideas about right and wrong" (Singer 2005: 349). Some theorists have argued that human beings have evolved a special faculty of moral judgment that functions in a way parallel to the way innate syntactical structures function in governing our use of language. Colin McGinn attributes this conception of moral knowledge to Noam Chomsky, who originally formulated the parallel account of our knowledge of language. McGinn writes:

> According to Chomsky, it is plausible to see our ethical faculty as analogous to our language faculty: we acquire ethical knowledge with very little explicit instruction, without great intellectual labour, and the end-result is remarkably uniform given the variety of ethical input we receive. The environment serves merely to trigger and specialise an innate schematism. Thus the ethical systems of different cultures or epochs are plausibly seen as analogous to the different languages people speak – an underlying universal structure gets differentiated into specific cultural products.
>
> (1993: 30)

Common moral intuitions are the deliverances of this biologically innate moral faculty. (For a recent, elaborately detailed working out of this idea, see Mikhail 2011.)

One concern about this hypothesis, however, is that there seems to be no reason to suppose that natural selection would have produced in us a faculty of judgment that would track moral truth rather than reproductive advantage. Another possibility, of course, is that the psychological capacities we have evolved for other purposes enable us to reason not only about matters relevant to survival and reproduction but also about moral matters, and the substantial uniformity of moral intuition over time and across cultures is the result of our applying the same capacities to the same object – namely, moral reality.

Some philosophers, however, have argued that it is explanatorily superfluous to suppose that our intuitions are responses to an independent moral reality when evolutionary theory has already supplied a plausible account of their origin in the mechanisms of natural selection. It seems uncontroversial, for example, that common intuitions about parental responsibility, marital fidelity, loyalty, promise-keeping, and so on have an evolutionary basis. Evolutionary biologists and psychologists have produced elegant explanations of various forms of moral behavior by reference to such notions as inclusive fitness and reciprocal altruism, and one can be confident that more and more instances of moral behavior and belief will come within the scope of biological explanation as our understanding advances. It may therefore seem more plausible to regard our intuitions as the products of natural selection than as glimpses of moral reality.

Yet many people have certain intuitions that are contrary to those that evolutionary theory would predict they would have. It is difficult, for example, to find any reproductive advantage in the conviction of early abolitionists that slavery was

wrong, or in the intuition that it is in general wrong to harm or kill nonhuman animals, or in the belief that we have exacting duties to aid impoverished people in other countries, to reduce our own fertility to control population growth, and to make other sacrifices in our own quality of life to avoid adversely affecting that of future people. Generally such beliefs are also, at least initially, at variance not only with individual self-interest but also with the received beliefs, including the religious beliefs, within the culture of those who have them. Yet these beliefs often spread, though slowly and more by force of example than through persuasive reasoning. They are, moreover, often genuine intuitions rather than the products of inferential reasoning – as, for example, in the case of a child who, intuitively repulsed at the discovery of what meat really is and how it is produced, is prompted to demand to be allowed to become a vegetarian.

That people have such beliefs, and that in time they often come to be widely regarded as evidence of moral progress, supports a conception of moral intuition that is compatible with moral realism; namely, that some intuitions – those that survive both the initial process of filtering and the testing for consistency – are judgments that are true and that direct us to the discovery of foundational moral principles that are also true.

Acknowledgments

I have incurred various debts in writing this article: to David Boonin, Walter Feinberg, and Hugh LaFollette for comments on the earliest draft, to Ingmar Persson for comments that assisted my revisions for this second edition, and to unpublished work by Regina Rini, and especially to unpublished work by William FitzPatrick, for ideas that shaped some of my arguments in the last section "Challenges."

References

Barzun, Jacques (1983) *A Stroll with William James*, New York: Harper & Row.

Brandt, Richard B. (1979) *A Theory of the Right and the Good*, Oxford: Clarendon Press.

Gaus, Gerald (1996) *Justificatory Liberalism: An Essay on Epistemology and Political Theory*, New York: Oxford University Press.

Hare, R.M. (1975) "Abortion and the Golden Rule," *Philosophy and Public Affairs* 4: 201–22.

Huemer, Michael (2005) *Ethical Intuitionism*, Basingstoke, UK: Palgrave Macmillan.

McGinn, Colin (1993) "In and Out of the Mind," *London Review of Books* (2 December): 30–1.

Mikhail, John (2011) *Elements of Moral Cognition: Rawls' Linguistic Analogy and the Cognitive Science of Moral and Legal Judgment*, Cambridge: Cambridge University Press.

Rachels, James (1986) *The End of Life: Euthanasia and Morality*, Oxford: Oxford University Press.
Rawls, John (1972) *A Theory of Justice*, Oxford: Clarendon Press.
Ross, W.D. (1930) *The Right and the Good*, Oxford: Clarendon Press.
Ryan, Alan (2011) "The Passionate Hero, Then and Now," *New York Review of Books* (December 8): 59–63.
Sayre-McCord, Geoffrey (1996) "Coherentist Epistemology and Moral Theory," in *Moral Knowledge?* eds. Walter Sinnott-Armstrong and Mark Timmons, New York: Oxford University Press, pp. 137–89.
Sidgwick, Henry (1907) *The Methods of Ethics*, 7th edn, London: Macmillan.
Singer, Peter (1974) "Sidgwick and Reflective Equilibrium," *The Monist* 58: 490–517.
Singer, Peter (1993) *Practical Ethics*, 2nd edn, Cambridge: Cambridge University Press.
Singer, Peter (2005) "Ethics and Intuitions," *Journal of Ethics* 9: 331–52.
Timmons, Mark (1987) "Foundationalism and the Structure of Ethical Justification," *Ethics* 97: 595–609.
Unger, Peter (1996) *Living High and Letting Die: Our Illusion of Innocence*, New York: Oxford University Press.

Further Reading

Berker, Selim (2009) "The Normative Insignificance of Neuroscience," *Philosophy and Public Affairs* 37: 293–329.
Daniels, Norman (1996) *Justice and Justification: Reflective Equilibrium in Theory and Practice*, Cambridge: Cambridge University Press.
DePaul, Michael R. (1993) *Balance and Refinement: Beyond Coherence Methods of Moral Inquiry*, London: Routledge.
Greene, Joshua D. (2008) "The Secret Joke of Kant's Soul," in *Moral Psychology, Volume 3: The Neuroscience of Morality: Emotional, Brain Disorders, and Development*, ed. Walter Sinnott-Armstrong, Cambridge, MA: MIT Press, pp. 35–79.
Lillehammer, Hallvard (2011) "The Epistemology of Ethical Intuitions," *Philosophy* 86: 175–200.
Street, Sharon (2006) "A Darwinian Dilemma for Realist Theories of Value," *Philosophical Studies* 127: 109–66.

Part II

Factual Background of Ethics

Chapter 6

Ethics and Evolution

Richard Joyce

Evolutionary Ethics

The field known as *evolutionary ethics* has burgeoned over the last couple of decades, yet its remit and defining concerns remain obscure. In terms of both conceptual clarification and labor, evolutionary ethics can be usefully divided into an empirical and a philosophical program. I will provide a quick overview of both and then discuss each in more detail.

Empirical Evolutionary Ethics

Humans are (perhaps uniquely) moral creatures. In this context, the statement does not mean that humans are morally praiseworthy or admirable (or, for that matter, blameworthy or iniquitous); nor does it mean that humans are proper subjects of moral concern. Rather, it means that humans make moral judgments: we classify our world in terms of moral values, our actions in terms of moral rules, our character traits in terms of moral virtues and vices, and so forth. Where does this way of thinking come from? One may pose this question synchronically, and it is the job of moral psychology to answer that question – to reveal what faculties are involved in moral judgment. But one may also ask the question diachronically, in which case one investigates by what processes humans came to have those faculties involved in moral judgment in the first place. One possibility is that moral judgment is a relatively recent cultural invention that exploits various psychological faculties which evolved for other purposes. Another possibility is that there exist in the human mind mechanisms that evolved specifically to make moral judgment

The Blackwell Guide to Ethical Theory, Second Edition. Edited by Hugh LaFollette and Ingmar Persson.

possible. On the latter hypothesis – which can be called "moral nativism" – a faculty for making moral judgments is a biological adaptation that emerged because this way of thinking provided our ancestors with some sort of reproductive advantage over their competition.

Moral nativism is an empirical matter. Of course, there may be concepts that require clarification prior to embarking on the investigation (most obviously, what is a *moral judgment*), so a certain amount of a priori examination is involved in the discussion. But, ideally, once these clarifications from the armchair have been made, then the matter can be turned over to scientists. Part of what makes empirical evolutionary ethics exciting is the wide range of scientists whose skills are relevant to the hypothesis. Key contributions can be made by social and developmental psychologists, experimental economists, neuroscientists, geneticists, primatologists, anthropologists, comparative ethologists, and evolutionary biologists. A terminological oddity of the field is that although explicit discussion of moral nativism from any such scientist can be placed under the rubric "evolutionary ethics," this empirical enterprise is not a field of *ethics* in the traditional sense. Thus, an evolutionary psychologist, say, who advocates moral nativism, will often be taken to count as an evolutionary ethicist but not as an ethicist. (Proponents of moral nativism include Alexander 1987; Krebs 2005; Dwyer 2006; Hauser 2006; Joyce 2006; Mikhail 2011; Kitcher 2011.)

Philosophical Evolutionary Ethics

By contrast, philosophical evolutionary ethics *is* a subfield of ethics. The philosophical evolutionary ethicist proposes that facts about human evolution can help address certain perennial problems in moral philosophy. Moral nativism (if true) represents a prominent example of the kind of "fact about human evolution" that may have some bearing on moral philosophy. But even if moral nativism is rejected, there are other possibilities for philosophical evolutionary ethics. For example, the uncontroversial fact that humans have evolved by natural selection to be social creatures (even if lacking an innate moral faculty) might be appealed to in order to support or undermine certain ethical theses.

Some have given in to the temptation to call philosophical evolutionary ethics "*normative* evolutionary ethics," but this is misleading and it is instructive to note why. Moral philosophy is a wide-ranging and varied field, the outputs of only some areas of which should be thought of as *normative*. At one extreme lies applied ethics, which may offer definite practical advice on how to act in concrete scenarios (e.g., regarding euthanasia). Normative ethics is the enterprise of building a general theory of moral action that is applicable across all or a large range of cases (e.g., utilitarianism). At the other end of the spectrum lies metaethics, which is concerned with a number of interrelated theoretical matters, such as the ontology of moral properties, the nature of moral language, and the epistemological status of moral judgments. Philosophical evolutionary ethics is the investigation

of whether facts about human evolution might have some nontrivial constructive input to moral philosophy at any point on this vague spectrum. But this need not be a *normative* impact. Consider, for example, the long-standing metaethical debate over whether moral language functions assertorically or whether it performs some other speech act. Were some finding about the nature of human evolution to help settle this dispute – demonstrating, say, that moral utterances are assertions – this would hardly be a *normative* result.

This is important to note because philosophical evolutionary ethics is sometimes rejected out of hand on the grounds that it must at some point purport to derive normative claims from descriptive premises – a move that allegedly commits a logical fallacy (deriving an "ought" from an "is"). But this is not so. For example, the aforementioned metaethical thesis – "Moral judgments are assertions" – is not a normative claim. A claim about how the word "ought" functions in everyday speech is not itself an "ought"-claim. Moreover, even when the philosophical evolutionary ethicist does maintain that evolutionary considerations lend support to some normative claim or other, it is unlikely that he or she thinks that the descriptive (evolutionary) premises *alone* imply that result. And there is nothing fishy about drawing an "ought"-conclusion from premises, some of which are descriptive and some of which also contain "ought"-statements. In short, the common accusation that philosophical evolutionary ethics must fail because its advocates must be endeavoring to derive an "ought" from an "is" usually misidentifies what the philosophical evolutionary ethicist is attempting.

Philosophical evolutionary ethics can be divided into two opposed programs: vindicating and debunking. There are different kinds of vindication possible. One might seek to use evolutionary data to establish a positive result in applied ethics (e.g., "An act of euthanasia in such-and-such circumstances would be wrong"), or a result in normative ethics (e.g., "Utilitarianism is the best moral theory"), or in metaethics (e.g., "Moral properties objectively exist"). But not any positive metaethical result will count as a vindication because some metaethical theses are considered to be, by their very nature, antithetical to morality. For example, the error theoretic view – that moral judgments aim at the truth but systematically fail to secure it (see Mackie 1977) – is for obvious reasons considered as hostile to morality as a practice. Thus if one could use evolutionary data to help establish the error theory, then although this would in some sense "vindicate" a certain metaethical view, it would be more natural to consider it as a form of evolutionary *debunking*. (The same could be said of the epistemological thesis that all moral judgments are unjustified.) An evolutionary ethicist may also try to use evolutionary data to debunk one view while vindicating another. Edward O. Wilson, for example, in collaboration with Michael Ruse, writes in debunking spirit that Darwinism shows us that "ethics as we understand it is an illusion fobbed off on us by our genes to get us to cooperate" (Ruse and Wilson 1985: 52); and yet elsewhere Wilson opines that a proper evolutionary understanding of ethics will "make possible the selection of a more deeply understood and enduring code of moral values" (Wilson 1978: 196). Peter Singer (2005) and Joshua

Greene (2003) use evolutionary considerations to debunk certain deontological normative theories, maintaining that this provides support for some version of consequentialism.

A sketch of the bipartite field of evolutionary ethics having now been provided, the rest of this essay basically repeats this pattern but in more detail. The treatment will be asymmetric, expending more than three times the ink on the former task than the latter. No particular view will be defended; my aspirations do not extend beyond providing an overview (though, inevitably, a somewhat biased one) of the dialectic.

Moral Nativism

One should not really speak of "moral nativism" in the singular, for there are numerous hypotheses deserving of the title. While moral nativism is usually identified with the thesis that human morality is innate, there are two glaring sources of indeterminacy: (1) What does "innate" mean in this context? (2) Which trait(s) does "human morality" denote?

What Is Innateness?

Several commentators have noticed that different ideas are uncomfortably gathered under the term "innate" (Griffiths 2002; Mameli and Bateson 2007). One idea privileges the notion of a trait whose developmental emergence is relatively insensitive to environmental variation. Another idea focuses on a trait's being essential to species membership. Another important use of "innate" denotes a trait that has been selected for by natural selection. All three uses would need further tightening before they could be treated as scientifically respectable, but even these loose versions suffice to reveal nonequivalency. Cats' having four legs is both developmentally robust and a feline adaptation, but is not an essential feature of being a cat. The genetic impairment Down's syndrome is developmentally robust but is not an adaptation. Singing behavior is an adaptation for certain birds but sometimes requires crucial environmental input. And so forth.

When it comes to debates over moral nativism, different writers often use different conceptions of "innate," and even when they are explicit as to what they intend, the overall consequence is a degree of confusion. For example, in his paper "Moral Nativism: A Sceptical Response," Kim Sterelny is careful to explain that he is skeptical of the *developmental* nativist thesis: he allows that "there is a plausible . . . case for the idea that moral cognition is an adaptation," but adds that "even if that is right, it does not follow that this capacity is innate" (2010: 280). If such qualifying comments are overlooked, however, then one might gain the impression that Sterelny is opposed to those who advocate moral nativism as an

adaptational hypothesis, when in fact they may be in complete agreement. It is therefore vital to keep in mind what sense of innateness is under discussion at any given time.

In the present context, focused on the relation between evolution and ethics, it is some version of the adaptational hypothesis that is under consideration. That is not to say that the developmental version of moral nativism is irrelevant – its truth or falsity may have important bearing on the evolutionary thesis – but it is the adaptational hypothesis that is our target. But this does not quite settle what is meant by "innate" here, for one may wonder what kind of *adaptation* is intended. Adaptations are traits that emerge and persist in a population through a process of selection: they are transmitted from parents to offspring and provide their bearers with a reproductive advantage. But this central Darwinian idea is neutral concerning the mechanisms of inheritance. The widespread assumption that genetic transmission is the only mechanism that counts is mistaken; traits passed on through cultural channels can just as well count (by orthodox Darwinian standards) as adaptations (see Boyd and Richerson 1985; Richerson and Boyd 2005). Hence, one could legitimately claim that human morality is a Darwinian adaptation (thus advocating moral nativism) while also maintaining morality to be an entirely cultural phenomenon. Nevertheless, this is not the way the matter is generally seen at present. It seems fair to say that nearly everyone who discusses moral nativism as an evolutionary hypothesis has in mind the view that the core of morality is a *biological* adaptation transmitted via genetic mechanisms, and this is what I shall assume in what follows.

Let us now turn to the nature of the trait, which I have been calling simply "human morality." The earlier introductory outline specified that this phrase denotes the capacity to make moral judgments. But further indeterminacy arises, since it is not at all clear what makes a judgment a *moral* judgment, and even if that were settled there remain a number of distinct hypotheses concerning what aspect of moral judgment might be an adaptation.

What Is a Moral Judgment?

Let us start by distinguishing moral nativism from what might be called "altruism nativism." Here I use "altruism" in the vernacular sense, to denote certain motivations and actions performed from those motivations. An altruistic act is one that is done with the ultimate goal of benefiting another. Even if such an action ended up harming the intended beneficiary, we might continue to call it "altruistic" (though perhaps appending "misguided"). An act that is done in order to benefit another but only because in providing that benefit one hopes to profit *oneself*, is not an altruistic act. Since altruism in this sense requires the cognitive capacity to conceive of *oneself* and of *others*, it (and its contrasting sense of *selfishness*) can be sensibly ascribed only to cognitively sophisticated creatures. Indeed, humans are the only undisputable case.

It is important to distinguish this psychological sense of altruism from *evolutionary altruism*. The latter is a characteristic of traits that have been selected for because they decrease the reproductive fitness of their bearer while increasing the fitness of others. Evolutionary altruism has nothing to do with psychological states; it is something that can sensibly be ascribed to plants or protozoa. Many cooperative behaviors in nature (social grooming, food sharing, group hunting, etc.) are casually referred to as instances of "evolutionary altruism" when in fact this is questionable. It depends on whether one understands "fitness" to mean (1) fitness over the life of the organism, and (2) reproduction that is achieved indirectly as well as directly. A primate that grooms another because the favor will be repaid in the future is not sacrificing its fitness if we measure fitness over its life time. A wild turkey that foregoes seeking mates in order to help its brother in his mating displays is not sacrificing its fitness if we measure fitness in a way that includes its indirect (inclusive) contributions to the replication of its genes. With these restrictions made, one may wonder whether genuine evolutionary altruism is even a possibility. Elliot Sober and David Sloan Wilson (1998) argue that it is a possibility (and, indeed, a reality) so long as populations are structured in such a way that the altruistic trait, though detracting from the reproductive fitness of its bearer relative to other group members, nevertheless provides reproductive advantage to that group relative to other groups.

Cooperation is often the means to reproductive success. When natural selection favors cooperation – whether it is evolutionarily altruistic or evolutionarily selfish – then proximate mechanisms governing the behavior will be necessary. These mechanisms will take a myriad of forms in nature, for they include the workings of everything from algae to zebras. For creatures with the cognitive wherewithal, there is no reason why natural selection may not plump for *altruism*, in the aforementioned psychological sense, as a capacity to encourage forms of cooperation. For example, a parent who simply cares directly for the welfare of his or her children is, arguably, moved by a more reliable and frugal mechanism than the psychologically selfish parent who must perceive a contribution that the children's welfare makes to his or her own well-being (see Sober 2000; see Stich 2007 for criticism). Though the argument seems strongest when applied to parent–offspring bonds, there is nothing to prevent its being employed in reference to other kinds of cooperative relations that were adaptive in the ancestral environment, such as reciprocal and mutually beneficial interactions among nonkin.

An important point to underline is that the hypothesis that such genuinely altruistic traits have evolved in humans is entirely consistent with the processes that produced them being evolutionarily selfish. Or, to put it another way, that a cooperative trait turns out to be evolutionarily selfish in no way shows that the cooperative behavior is "really" psychologically selfish. This goes even for reciprocity. In a reciprocal exchange, the proportionally greater payback one receives for one's efforts may explain why the capacities to engage in such behavior were selected for, but it is an error to assume that this reveals anything about one's psychological motivations. Of course, in humans reciprocal relations are often

governed by pragmatic thoughts of payback, but they need not be; there is no a priori reason for excluding the possibility of evolution preferring to encourage such behavior by granting parties (possibly conditional) altruistic concerns for their exchange partners.

Such considerations form a plausible basis for supposing that nativism concerning human psychological altruism might be true. But considerations in favor of altruism nativism fall short of considerations in favor of moral nativism. To see why, one need only reflect on the fact that a person can be moved by altruistic sentiment while making no moral judgment whatsoever. A parent's powerful motivation to care for his or her children is typically governed simply by the raw *desire* to do so, not by any moral consideration. This is not to deny that such a parent's actions are morally praiseworthy, and nor is it to deny that parents would acknowledge the presence of parental duties if asked. Rather, the claim is that parents often have entrenched and reliable feelings of affection for their children independently of any moral judgments. Acting in a helpful manner because one wants to is simply not the same as acting in a helpful manner because one judges that one ought to. It seems that we can, in fact, imagine a social species that evolves with strong feelings of altruistic concern for each other but whose idiosyncratic neurology renders them entirely lacking in the conceptual prerequisites necessary for making moral judgments. Such creatures cooperate because they love each other, but they have no capacity to think of cooperation as a morally desirable or morally required activity. Altruism nativism would be true of such creatures, but moral nativism would not be. (Again, we may decide that such creatures are worthy of our moral praise, but that is not the current concern.)

Perhaps such imaginary creatures are not a million miles away from some of our primate cousins. Primatologist Frans de Waal has argued that apes and some monkeys have many of the "building blocks" of morality, such as empathy, an awareness of equity, and the capacity to engage in consolation behavior (1992, 1996, 2006). But while these traits might suffice for what de Waal calls "a sense of social regularity" (1992: 242), he admits that they fall short of anything truly deserving the name "a moral sense." But what other building blocks need to be added? What is it that the chimpanzee lacks?

A tempting (though contentious) answer is that the chimpanzee lacks certain *cognitive* resources. A chimp that uses its canine teeth during an in-group dispute incurs an angry response from the other group members, even those not directly involved in the conflict (de Waal 1992: 247). Yet it is unlikely that they think of the action as *justifying* the reprimand; it is doubtful that they can conceive of the action as wrong in a coolheaded way, in the absence of angry arousal. The intensity of the punitive response will depend entirely on the contingent intensity of the immediate arousal; it will not derive from deliberation concerning what level of reprimand is *fair*. The transgressing chimp may fear the negative response from the group, but it seems improbable that he can grasp that he might *deserve* this response – a conceptual achievement necessary if the emotion of *guilt* is to be ascribed.

These thoughts are little more than gestures in a certain direction, but they do at least accomplish that: to point toward a promising line of thinking. According to this line of thinking, moral judgment requires the capacity to deploy concepts like *desert*, *wrongness*, and *justification*. An additional common idea is that moral judgments are distinguished from other kinds of normative judgment (such as concerning prudential and conventional norms) by their being imbued with a certain distinctive practical authority. Moral prescriptions are those that one *has to* follow whether one likes it or not; moral imperatives, in other words, are not pieces of advice on how to satisfy one's ends (see Kant 1783/1985; Joyce 2001). If the practical authority of moral imperatives does not derive from their conducing to one's ends, then whence does it derive? Arguably not from any human authority. Moral norms are often contrasted with conventional norms on the grounds that the latter are allowed to depend on the decree of some authoritative body (be it monarchs, teachers, God, or collective opinion), whereas the former are taken to be transcendent of any institution (see Smetana 1993; Nucci 2001; Turiel 2002). (Note that we are discussing the authority with which moral norms are imbued; whether moral norms actually *have* this authority is another matter entirely.)

The preceding is admittedly all rather vague and is certainly controversial. This is one of the difficulties with the debate over moral nativism: people argue at length over the origins of a trait but often have only a slippery grasp of what the trait is, or, even when they have a crisp idea in mind, their stipulative construal of the trait differs from that of others. Until there is agreement as to which trait is under discussion, confident claims about its historical source are rather premature – an admonition applying as much to the antinativist as to the nativist. Nevertheless, with this important hesitation noted, there is nothing to stop one signing a promissory note to attempt to settle the matter in the future, and proceeding further with the discussion in a provisional fashion.

Different Moral Nativisms

Even if we knew precisely what a moral judgment is (or at least made a collective decision in this regard), there would still be some question as to what moral nativism is. At one extreme is the nativist claim that no particular moral judgment is innate but that the general faculty for making these judgments is innate. One way of conceiving of this view is that the human brain comes prewired with moral *concepts* (concepts like *wrongness*, *goodness*, etc.) and the social environment teaches the individual to which items these concepts should be attached. Thus one environment may teach the child to judge that killing foreigners is morally wrong, while another environment may teach another child to judge such behavior as acceptable or even admirable. This variation would be no mark against the kind of nativism under discussion, for both parties would equally exhibit the trait in question: the tendency to make moral judgments.

Another kind of moral nativism holds that the concepts provided by the moral faculty come with biases toward certain subject matters. Psychologists Jon Haidt and Craig Joseph (2004), for example, argue that a comprehensive survey of cross-cultural data reveals that the moral sense comes prepared to manage certain broad domains: actions producing harm, relations concerning fairness and exchanges, values pertaining to social hierarchy, and regulations surrounding certain bodily matters (such as menstruation, sex, corpses, etc.). According to this view, humans will find certain moral systems easier to learn than other systems. (See also Haidt and Bjorkland 2008.)

Another kind of nativism will hold that specific moral judgments are biological adaptations. Clearly there is nothing to be said for the view that *all* moral judgments are biological adaptations (for consider a moral judgment about that particular lie that Fred told last Tuesday). But perhaps for certain general complete judgments, concerning matters that were present in the prehistoric environment, the nativist claim might be upheld. Jesse Prinz (2009), for example, discusses nativism for the judgments "Don't harm innocent people," "Respect and obey authorities," and "Incest is prohibited." (Prinz argues emphatically against this nativism, but the fact that he finds it necessary to do so shows that it must be a viable hypothesis.)

Perhaps a few broad abstract moral principles are prewired into the human brain, and in addition there are certain innate parameters, such that the social environment sets with which parameter the principles combine in order to produce a full range of specific moral judgments. This is the position of John Mikhail (2011) and Marc Hauser (2006), drawing inspiration from Noam Chomsky's view on linguistic nativism.

This significant variation in different forms of moral nativism is important to appreciate; since some versions may be more plausible than others, evidence mustered in support of one version may not serve to support others, and sound arguments against one version may not trouble others.

Why Might Moral Thinking Have Been Adaptive?

Any evolutionary moral nativist owes us an account of why moral judgment (however it is to be defined) was reproductively useful – what the "adaptive problem" was to which the trait was the solution. Such an account need not be pure speculation; it can to some extent be based on careful empirical observation of the ways in which moral thinking continues to impact upon our practical lives. The provision of such an account does not show that nativism is true, of course, but nevertheless contributes to its plausibility.

On this matter (as with every other element of moral nativism we have thus far discussed) there are various options for the nativist. One important distinction concerns *whose* reproductive fitness is enhanced. It is natural to assume that one should examine the profits for the individual who bears the moral trait. If, however,

one is inclined to suspect that the moral faculty evolved via a process of group selection – of the kind associated with Sober and Wilson's views on evolutionary altruism – then it is the reproductively relevant profits to the bearer's *group* that are pertinent. Another distinction concerns whether one is considering *self-oriented* or *other-oriented* moral judgments. It is possible that one kind of moral judgment was primarily adaptive and the other emerged subsequently as a kind of useful by-product. Darwin, for example, in *The Descent of Man* frequently identifies the moral sense with the *conscience*, indicating a focus on self-directed moral judgments, whereas Edvard Westermarck offers the reverse view: that self-oriented moral assessment is attained only "through a prior critique upon our fellow-men" (1906: 123).

Whatever options are taken in these respects, most moral nativists suggest that moral judgment was adaptive because it in some manner encouraged ancestral cooperation. It is plausible that in certain contexts thinking of a cooperative venture as, say, *obligatory* – such that defecting on the deal will leave one feeling that punishment is not merely risked but *deserved* – will strengthen one's resolve to perform the action (see Joyce 2006: ch. 4). Needless to say, this trait could be disastrous if attached to maladaptive behavior (e.g., over-cooperating), so the "moralization" will need to be sensitive to environmental variables. According to this view, the evolutionary function of morality is to act as a kind of motivation enhancer, to offset the possibilities of practical self-sabotage (e.g., succumbing to temptations of immediate gratification). These possibilities may arise because of the operations of other psychological faculties (e.g., our capacity to calculate for self-gain), which are generally adaptive but occasionally steer us toward suboptimal decisions.

It would be a mistake, however, to see the benefit of the moralization of certain cooperative behaviors simply in terms of a private little mental nudge. Such a view fails to appreciate the social dynamics of moral assessment. Moral evaluation is something that can be employed publicly to condemn an action or a person, to justify a punishment, and to demand others to participate in the punitive response. Moral values can be seen to be *shared*, thus enhancing social cohesion even when the moral value itself does not pertain to any cooperative behavior (but concerns, say, the treatment of food). Thus moral judgments can function usefully as personal commitments, but they can also be signaled in a way that makes them potential interpersonal commitments. The fact that abiding by moral standards generally involves foregoing short-term profits means that morality can function well as a *costly* signaling device. When choosing partners for a mutually beneficial cooperative venture, it makes sense to prefer those who can *honestly* signal their willingness to participate. And making signals costly is a way of making them honest, for a sufficiently expensive signal costs the signaler more than the profits that might be reaped through dishonesty (see Zahavi 1977; Noë 2001). Thus, if one's flourishing or very survival depends on being chosen in cooperative ventures (whether it be as a mate or as a member of a hunting party), it may be adaptive to signal in a costly way one's social virtues.

In accounting for the adaptiveness of moral judgments, the moral nativist may therefore choose to downplay their role as motivation enhancers and instead highlight their potential as grounding public signals. Psychologist Geoffrey Miller (2000, 2007), for example, explores how morality serves as a signaling device in human mate selection, and thus argues for an unusual kind of moral nativism according to which the moral sense evolved through a process of sexual selection (see also Nesse 2007). Economist Robert Frank (1988) argues that moral emotions evolved as signaling devices – that the facial expressions accompanying emotions, which in large part lie outside our autonomous control, are nature's way of guaranteeing a degree of honesty in our social interactions in a manner that gave our ancestors who bore this trait a social advantage over the competition, which translated into a reproductive advantage.

It is not necessary to see the adaptiveness of moral judgment in terms of its social role. Daniel Dennett (1995) suggests that moral thinking serves as a "conversation stopper": a no-questions-asked practical consideration that silences any further deliberation on the matter in question. In certain circumstances, we humans are too clever for our own good; we have the capacity to keep worrying about costs and benefits, possible consequences, and so on, to such an extent that the very labor of calculating the optimal course of action ensures a suboptimal outcome. Hence, absolute categorical rules and values can play a very useful role in deliberations, bringing them to an endpoint on an outcome that is probably satisfactory. There is no need, however, to assume that these rules and values must concern social matters (like *Do not kill your neighbors*); they may fulfill this function just as well if they concern entirely private affairs (like *Do not risk your own life for trivial gains*).

While the moral nativist certainly needs to say something about how moral thinking was adaptive for our ancestors, in one sense this is the least challenging aspect of making the case, since nearly everyone – antinativists included – accepts that moral thinking serves some useful social purposes. Of course, moral thinking can lead to disasters as well, so there is a substantive question as to whether it is all-things-considered a good thing. (For doubters, see Hinckfuss 1987; Garner 2010; and, in certain moods, Nietzsche 1887/1994.) But the moral nativist can expect to gain agreement from at least most of the opposition that morality was and is generally useful (if not to the bearer of the trait, then to his/her group), and this usefulness is likely to translate in some manner into a plausible case for reproductive advantage. The harder challenge for the moral nativist is to find evidence for the plausible hypothesis.

Evidence for Moral Nativism

Perhaps the only claim that can be made confidently about the status of evidence either for or against moral nativism is that the whole debate is in a state of disarray. There is a fundamental lack of consensus concerning what would even count,

in general, as evidence that a psychological trait is an adaptation; and therefore it comes as no surprise that there is much confusion when it comes to the trait of moral judgment, about which, as we have seen, there is no fixed view as to what the trait even is. (Discussions of empirical methodology regarding psychological adaptations include Ketelaar and Ellis 2000; Conway and Schaller 2002; Confer *et al.* 2010.)

Once a version of the moral nativist hypothesis has been focused upon, testing must proceed as with any empirical hypothesis: predictions of the hypothesis must be identified, the evidence of whether these predictions obtain must be gathered, and alternative explanations of the data must be examined. While such truisms can be agreed upon, when it comes to testing an adaptationist hypothesis there is trouble at every step.

It is often assumed, for example, that one of the predictions of nativism is the universal (or near-universal) presence of the trait in human populations. Thus, much energy is expended on cross-cultural studies, both by nativists trying to demonstrate universality and by antinativists trying to provide counterexamples. If the hypothesis under discussion is nativism *in the developmental sense* mentioned earlier (see "What is Innateness"), then evidence pertaining to universality can be effective, for moral nativism in that sense is the claim that moral judgment is a trait whose emergence is reliable in the face of environmental perturbation; thus to observe morality emerge ubiquitously in a highly varied range of developmental settings would constitute solid confirming evidence. But the relation between *evolutionary* nativism and universality is much more complicated, to say the least.

It was noted earlier that adaptations may require specific environmental input. This may be in the form of external "cues" that trigger development. If these cues were reliably present during the period of selection, then there would be no pressure to make the emergence of the adaptive trait assured in the absence of the input (see Griffiths 2002: 74–5). Yet the modern environment differs in countless ways from the ancestral one, raising the possibility that certain cues are now given unreliably, in distorted form, too often, or not at all. Alternatively, the developmental process may be "expecting" differential cues, signaling the need to take one developmental trajectory or another. For example, whether male European earwigs (*F. auricularia*) develop long or short forceps, which determine distinct mating strategies (the two forms being so different that they were once considered distinct species), depends on local population density (Tomkins and Brown 2004). In the latter case, what is selected for is a mechanism that executes a conditional strategy depending on environmental variables. In the former case, even the conditional mechanism might not develop due to the absence of crucial environmental cues during development (see Buller 2006: ch. 2). In either case, we should not expect to see a universal phenotype associated with the adaptation.

The reverse implication also fails. Even if we do find a universally present trait, it is very difficult to exclude alternative nonadaptationist hypotheses. The trait might not be itself an adaptation but rather a "spandrel" that piggybacks on other adaptations. Male nipples are universally found (in males) yet are not an adaptation

but a spandrel (Gould 1992). In the case of moral judgment, a conspicuous alternative hypothesis is that moral judgment serves no evolutionary function but is an inevitable accompaniment of other psychological faculties that are adaptations. If these other adaptations enjoy universal manifestation, then so too will the moral capacity. Alternatively, even if moral judgment does not *inevitably* accompany adaptational traits, these traits might make moral judgment possible for the human mind, and moral systems might simply be a fairly obvious invention for groups of humans living together. This is what Dennett calls a "Good Trick": a solution to a recurrent problem that is sufficiently straightforward for creatures with our array of sophisticated cognitive and emotional capacities that one can expect it to be struck upon pretty much everywhere (Dennett 1995: 77–8, 485–7). The observation of universality alone does not favor the nativist hypothesis over either the spandrel hypothesis or Good Trick hypothesis.

The poverty of the stimulus (POS) argument is another kind of evidence in favor of nativism that is more problematic than is often realized. The structure of the argument comes, of course, from the debate over nativist explanations of human linguistic abilities (see Chomsky 1967, 1987/1990), where the POS argument is widely judged to be triumphant in establishing some form of nativism. According to this argument, the capacities evident in language use emerge in a manner that far outstrips the information that is available in the learning environment. What are the prospects of a *moral* POS? The answer is that there are challenges at every step of the argument.

First of all, it is not clear to what extent the hypothesis that moral judgment is a biological adaptation predicts that the trait will emerge in conditions of impoverished stimuli. Some versions of moral nativism described earlier require a substantial amount of *learning* in the acquisition process. According to one version, for example, what biological selection provides is a mechanism that makes possible a particular special kind of learning (i.e., the acquisition of moral norms). If such a version is under scrutiny, it would be a distraction to examine how a child acquires a full moral judgment, like *stealing toys is wrong*, for we can all agree that the child is taught this by his/her parents. Rather, what one should be asking is how the child comes by the concept *wrongness* in the first place. But even here, as we have seen, the hypothesis that the possession of this concept is a biological adaptation is consistent with its developmental emergence requiring specific environmental input.

Second, it is debatable to what extent the observable data reveal that the moral acquisition process does unfold in conditions of poor stimuli. A lot of research is understandably focused on how *early* traits appear in development. Psychologist Paul Bloom and colleagues have found evidence that infants as young as three months preferentially discriminate others on the basis of their social behavior toward third parties (Hamlin, Wynn and Bloom 2007, 2010). Nobody maintains that these infants are making full-blown moral judgments, of course; but the full-blown trait in which we are interested appears surprisingly early. Following the lead of Elliot Turiel, several developmental psychologists have investigated the

emergence of the capacity to distinguish moral from conventional norms, finding evidence of the capacity in children yet to turn three (Smetana and Braeges 1990). Yet even here it is questionable to what extent this early emergence occurs in conditions of impoverished stimuli. By the age of three the child has been exposed to an enormous amount of unrelenting instruction from its parents, and has usually been able to observe a broad variety of social interactions. Shaun Nichols reminds us that "the child is exposed to lots of admonitions and instruction in the normative domain. Parents and teachers are constantly telling kids what shouldn't be done" (2005: 358). Sterelny makes a similar observation:

> The narrative life of a community – the stock of stories, songs, myths and tales to which children are exposed – is full of information about the actions to be admired and to be deplored. Young children's stories include many moral fables: stories of virtue, of right action and motivation rewarded; of vice punished. So their narrative world is richly populated with moral examples.
>
> (2010: 289)

When it comes to assessing how rich or poor is the child's learning environment, what is crucial (again) is identifying just which trait it is whose emergence is under discussion. But even making this decision leaves many challenges. Suppose, for example, that the target hypothesis concerns how a child acquires the basic moral concepts (*wrongness*, etc.) in the first place. The problem is that we do not know even in principle what combination of endogenous and exogenous factors are necessary or sufficient for such a concept to become available in a developing human brain. The puzzle can be put informally as follows: Let us grant the developing child all the careful instruction, all the scaffolded learning, all the varied experiences, all the trial-and-error social interactions, all the exposure to moral tales and exemplars, all the coordinated rewards and punishments, and so on, that one cares to – let us, in short, allow the stimulus to be as rich and varied as one likes; the challenge remains: what psychological mechanisms must be present in order for all this environmental input to result in the emergence of the capacity to employ a moral concept? Since there appears to be no consensus on how this question should be answered, the possibility remains open that one or more of the necessary mechanisms must be *dedicated* to this particular kind of acquisition, in the sense that they have been forged by biological selection for the task.

A third difficulty for a moral POS argument is that even if the adaptational hypothesis did predict that moral development would occur with impoverished stimuli, and even if we had solid evidence that moral development *does* have this characteristic, the moral nativist still faces the challenge of excluding alternative hypotheses, and this is no easy matter. It has often been noted that the conclusion of the POS argument is negative – it establishes one means (i.e., learning) by which the psychological capacities in question have not come about (see Laurence and Margolis 2001; Nichols 2005). The success of such an argument might support

nativism of a developmental sort, but it does not suffice for nativism in the evolutionary adaptational sense in which we are interested here. A trait may become manifest prior to any learning, indicating that it is in some sense "nonacquired," but this would not suffice to show that it is an adaptation. The prominent alternative hypothesis that is not excluded is that the trait is a spandrel. A spandrel that piggybacks on other traits, which are adaptations, will emerge just as the adaptations do. If the adaptational traits are expressed early, in a predictable sequence, universally, and in environments of impoverished stimuli, then so too will the spandrel trait.

Alternative By-product Hypotheses

Several opponents to moral antinativism have offered what amount to alternative spandrel hypotheses – or at least *by-product* hypotheses.[1] The standard approach is to offer some psychological faculties that are not specifically *moral*, then attempt to show that the capacity for moral judgment could (or would) emerge from these traits. Darwin himself can be interpreted in this way. He writes that "any animal whatever, endowed with well-marked social instincts, . . . would inevitably acquire a moral sense or conscience, as soon as its intellectual powers had become as well, or nearly as well developed, as in man" (1879/2004: 120–1). These "intellectual powers" include a good memory, language, and the capacity to form habits of thinking and acting. Thus it seems that the moral faculty, for Darwin, is a spandrel of other psychological capacities. Even regarding the social instincts themselves he is undecided as to whether they are adaptations or by-products, writing that it is "impossible to decide in many cases whether certain social instincts have been acquired through natural selection, or are the indirect result of other instincts and faculties" (1879/2004: 130).

Modern antinativists do not concur regarding Darwin's exact ingredients (the one who comes closest is Francisco Ayala (2010)), but the pattern of argument remains the same. Prinz (2008), for example, tries to build moral judgment out of nonmoral preferences (e.g., for fair reciprocal exchanges, against incest), metaemotions (emotions that take other emotions as their subject), perspective taking (allowing for third party concern), and nonmoral emotions such as sadness and anger. Nichols (2005) allows that nativism might be true of the capacity to use nonhypothetical imperatives (i.e., imperatives that do not depend for their legitimacy on the addressee having certain ends), but he rightly notes that moral imperatives are but a subset of nonhypothetical imperatives (e.g., "Do not speak with your mouth full" is a nonhypothetical but nonmoral imperative (see Foot 1972)). Nichols adds that this capacity might combine with another innate tendency – namely, an affective mechanism that responds to suffering in others – to explain why the domain of moral normativity gets singled out as salient, resonant, and memorable. He writes:

> Both of the mechanisms that I've suggested contribute to moral judgment might well be adaptations. However, it is distinctly less plausible that the capacity for core moral judgment itself is an adaptation. It's more likely that core moral judgment emerges as a kind of byproduct of (*inter alia*) the innate affective and innate rule comprehension mechanisms.
>
> (2005: 369)

Such offerings must be assessed on a case-by-case basis. The key question to examine is whether the nonmoral ingredients offered really do suffice to account for the capacity to make moral judgments. Simply asserting that they do is an easy undertaking, but a proper critical examination of the claim may be a complex matter requiring various kinds of empirical investigation. One factor which might confound the debate is the possibility that the very notion of *moral judgment* is to a greater or lesser extent indeterminate, such that the antinativist's ingredients suffice to explain moral judgment in some attenuated form while not accounting for a more full-blooded conception of the phenomenon (see Joyce 2013).

Any proponent of a by-product hypothesis needs to be careful that it really is a by-product hypothesis that is being articulated rather than a useful clarification of a nativist hypothesis. Let me explain using Nichols' suggestion as an illustration. On one reading (the one he intends), Nichols, if successful, will have accounted for the trait of moral judgment by appeal to other evolved mechanisms. But on an alternative interpretation, Nichols has simply succeeded in providing some extra detail for the nativist case. Notice that how we enumerate psychological "mechanisms" is always to some extent dependent on our focus; what from one theoretical perspective is a single mechanism is, from another perspective, a group of mechanisms, each of which is made up of a group of mechanisms. Thus, if humans do have an innate "moral faculty," one should not expect a monolithic entity that performs moral judgment (arguably not even a coherent idea), but rather a group of innate subfaculties. Perhaps, then, Nichols has simply succeeded in delineating two of those subfaculties.

The issue turns on whether these two mechanisms are "supposed" (from an evolutionary perspective) to work in tandem. On the one hand, we can picture two biologically evolved mechanisms ticking along, interacting in certain ways, and the consequent joint output having characteristics that are wholly accidental. In this case, the output should be considered a by-product or spandrel. On the other hand, we can picture this output having some reproductive relevance, and thus natural selection taking an interest in the interaction of the two mechanisms, tinkering with the human genome in order to encourage the interaction. In this case, at a certain point we must cease to speak of the output as a "by-product" of two innate mechanisms, and rather speak of it as issuing from a dedicated mechanism (which consists of two preadaptional submechanisms). According to this latter account, one might call the moral faculty an *exaptation* – a result entirely acceptable (and, indeed, unsurprising) to the moral nativist.[2]

Even if the moral antinativist succeeds in articulating a clear by-product hypothesis, and avoids the charge that what has really been accomplished is a delineation of the submechanisms comprising the adaptational moral faculty, this would not show that moral nativism is false. I know of no good reason for endorsing *methodological antinativism*, according to which if a nonnativist explanation of a trait is possible, then it is to be preferred. Such a principle is at best strictly optional, and at worst entirely question-begging. Of course, methodological nativism would be just as bad; but why presuppose "methodological *anything*-ism" in this respect? (George Williams once remarked that "adaptation is a special and onerous concept that should be used only where it is really necessary" (1966: 4), but even he admitted that this assumption really amounts to "doctrine.") Even if one can conceive of a way that general learning mechanisms and nonmoral faculties could suffice for moral judgment, the latter competency might nevertheless as a matter of fact be brought about by dedicated innate mechanisms. Natural selection might, for example, have reasons for favoring a more reliable and faster acquisition process over empirical learning, even if acquisition via learning remains possible. Empirical scrutiny of the actual acquisition process might reveal that the antinativist's plausible hypothesis is simply not how the process occurs.

In the foregoing I have not attempted to put forward evidence for or against moral nativism, but rather draw attention to the complexities of the matter. The fundamental message is that (to the extent to which the relevant concepts can be disambiguated and made precise) nativist hypotheses are orthodox empirical claims and must be treated as such. As with any area of science, there is a place for imaginative speculation and weighing matters in terms of what seems most plausible. Of course, if anyone thought that a plausible bit of speculation could be considered *true*, then Gould's famous allegation that this amounts to no more than a *just so story* would be fair. However, it is far from clear that any serious moral nativist does make so crass an error; most appear acutely aware that their hypotheses require evidential support and that the empirical project is still in its infancy.

Implications for Moral Philosophy

Ever since the ancient Greeks it has been possible to discern a broad dichotomy in moral philosophy: between those views that see the moral realm as somehow transcendental and objective and those views that see morality as dependent on embodied humans that are products of nature. Those who favor the latter perspective will find it hard to resist the thought that discoveries concerning the Darwinian processes that shaped our biological nature must have some relevance to human morality. But what kind of relevance? Possible answers are too numerous to be cataloged here.

The first philosopher to attempt to use Darwinian theory to justify particular moral decisions was Herbert Spencer. Spencer championed a progressive view of

evolution – that the inevitable outcome of natural selection is superiority of organisms (more complex, smarter, more cooperative, more harmonious): "The conduct to which we apply the name good, is the relatively *more evolved* conduct" (1879/2011: 26). From this dubious basis, he argued that certain laissez-faire social policies followed: that it is often the role of the state to let matters run their natural course with respect to society's "unhealthy, imbecile, slow, vacillating, faithless members" (Spencer 1851: 324). (In reality, this is but one thread of Spencer's changing and complex views on moral philosophy. See Weinstein 1998.) Arguably, Spencer's error here derives from confusion over what is meant by Darwinian *fitness* – hardly a surprise when it is recalled that Spencer coined the phrase "survival of the fittest": a problematic expression that has often been accused of tautological vacuity. Organisms that are *fitter* than their competitors in the Darwinian sense are simply those that replicate their genes more successfully. But these fitter organisms need not be fitter in the intuitive vernacular sense. The fitter may be less complex, less intelligent, and less cooperative; they may be more unhealthy and more prone to imbecility and faithlessness (if, for example, these are the traits strongly preferred by potential mates). Spencer slides between these different senses of "fitness," and thus fails to see that when the welfare state chooses to aid a starving pauper, say, it is not a case of an evolutionarily unfit organism being unnaturally propped up, but rather (in all probability) the organism in question being rendered evolutionarily fitter.

Contemporary philosophers have pursued various strategies for using evolutionary thinking to vindicate morality, either regarding particular practical decisions or more generally (see Richards 1986; Rottschaefer and Martinsen 1990; Dennett 1995). Here I will focus briefly on just one family of arguments that try to find moral normativity in the evolutionary process.

If moral judgment is an adaptation, then we can legitimately speak of its "usefulness," from which one might draw a conclusion concerning *justification* (Campbell 1996). Moral nativism also allows one to speak of morality's having a biological function, from which one might draw a conclusion concerning what morality is "supposed" to do, and of the moral sense operating "well" or "poorly" (Kitcher 2011). Such positive-sounding conclusions for morality are, however, not of the kind in which a moral philosopher is typically interested. From the fact that a trait is the product of biological natural selection we can draw two conclusions concerning its usefulness: first, that it *was* useful during the relevant period of selection; second, that it was useful for the replication of genetic material. It does not follow that it *is* useful now, nor that it is useful for any ends that an individual might have that do not concern genetic proliferation, of which surely we have an abundance. Perhaps even more important to note is that any such justification is entirely *instrumental* – not the kind of *epistemological* justification with which metaethicists are usually concerned.

From the fact that a trait's having a biological function gives license to an array of normative-seeming language (concerning what it is "supposed" to do, what is involved in its functioning "well," and so forth), no real practical authority arises.

By analogy with artifactual functions, if an archeologist unearths an ancient taro pounder she is under not a glimmer of obligation to use the item for pounding taro. Similarly, consider a character like Hume's sensible knave, who generally observes the norms of morality only on the grounds of expediency but who does not hesitate to take full advantage of any exception that arises, even if it is at great cost to others. It may well be that in choosing to break an inconvenient promise the knave is using his moral sense in a way that it is not "supposed" to be used, or perhaps failing to use it at all, but this fact alone – the fact that a biological function is going unfulfilled – does not seem sufficient to underwrite our assessment that he is acting in a morally prohibited manner.

This represents a serious problem for virtue ethicists who look to the ancient Greeks for inspiration but hope to appeal to Darwinian teleology to replace the bankrupt Aristotelian teleological worldview (see Casebeer 2003; Brown 2008). Evolutionary functions alone simply do not provide the action-guiding normativity with which virtue ethicists hope to imbue the virtues.

Recent years have seen explored a very different agenda concerning deriving metaethical conclusions from evolutionary data: the contention that moral nativism *undermines* morality. "Undermine" here is an intentionally vague term, covering different possibilities:

(1) Moral judgments are unjustified.
(2) Moral judgments are unjustified and irredeemably so.
(3) Moral objectivity is an illusion; therefore all objective moral judgments are false.
(4) All moral judgments are false.

Michael Ruse often couches his evolutionary moral skepticism in terms of lack of justification ("What kind of metaethical justification can one give for [moral] claims?" (Ruse 2006: 20)), but it would seem that his ultimate ambition is to establish a claim of falsehood ("Morality is a collective illusion foisted upon us by our genes" (Ruse 1986: 253)). Sometimes Ruse seems to be aiming to establish only thesis (3), but when one couples (3) with his explicit claims that objectivity is part of the *meaning* of morality – for example, "Ethics is subjective, but its meaning is objective" (Ruse 2006: 22; see also Ruse 2009: 507) – one can draw the conclusion that morality *tout court* is false. In other words, one might explicitly claim only that *moral objectivity* is a flawed notion, but if one also maintains that morality is necessarily objective (conceptually speaking), then one is in effect asserting that morality per se is flawed. (This, at least, is how I shall interpret Ruse here, if only to have a concrete advocate of thesis (4) before us.)

What role does moral nativism play in Ruse's debunking argument? Not only does he endorse moral nativism, but he argues that it explains why we imbue our moral claims with *objectivity*, since this is a crucial element in his account of why moral thinking was adaptive to our ancestors. Ruse then deploys a parsimony principle to establish skepticism. All that needs to be explained, he thinks, has been

explained. There is simply no need to go further and posit the realm of moral properties that is necessary to make moral judgments *true*. His favored analogy is with explaining the popularity of spiritualism after World War I. By appealing to psychological and sociological factors we can explain all we could want to about why a grieving mother might come to believe that her dead son spoke to her at a séance, without having to posit the realm of souls-talking-from-beyond-the-grave that would be necessary to make her belief *true*.

The analogy is not entirely persuasive, however. Whereas it is clear that what is needed to render the spiritualistic belief true is a whole layer of extra spooky ontology in the universe, this is not so obvious in the moral case. Many modern moral realists attempt to account for moral properties on the basis of antecedently accepted ontological categories. Hedonic utilitarianism, for example, proposes to build moral obligation out of happiness plus the causal relation of being productive of happiness. Thus it is not obvious that an application of Ockham's razor is apposite in the moral case.

A principle of parsimony also plays an important role in Sharon Street's attempt at an evolutionary undermining of morality (Street 2006). Her ambition is weaker than Ruse's, since she aims to use moral nativism to establish only thesis (3). Her opponent is moral realism (understood as including an endorsement of objectivity), and since she allows that some kind of moral constructivism might survive the argument, it is evident that she does not endorse the claim that objectivity is an essential part of the meaning of moral terms.

Street proceeds by observing that moral nativism presents the moral realist with a dilemma: Either (i) there is no causal connection between the (supposed) realm of objective moral facts and the moral beliefs with which natural selection has endowed us – in which case there is very little chance that our moral beliefs match the facts – or (ii) there *is* a causal connection, namely that our moral sense has been designed to track the objective moral facts. The problem with the second horn of the dilemma, according to Street, is that there is a superior hypothesis available – the "adaptive link account" – according to which moral judgment evolved because it encouraged our ancestors to behave in an adaptive manner in the social sphere. She writes that the latter is to be preferred on grounds of parsimony (inter alia), because "the tracking account obviously posits something extra that the adaptive link account does not, namely independent evaluative truths" (Street 2006: 129). As with Ruse's argument, however, one might complain that this appeal to parsimony has traction only if the "independent truths" in question involve positing ontological categories over and above those one antecedently accepts, and this is not obviously so for many forms of moral realism.

An evolutionary debunking argument with even weaker ambitions is one that I have myself developed (in Joyce 2006), which argues only for thesis (1). According to this view, the confirmation of a genealogical theory of moral judgment – one which neither implies nor presupposes that these judgments are true – would remove any justification which these judgments may have been accorded (on the grounds of epistemological conservatism, for example). Note that the conclusion

of the argument is not that any moral judgment is *false* (contra Mason's interpretation (2010)).[3] Note also that the purported *objectivity* of morality plays no role in the argument at this point (contra Kahane's interpretation (2011)); if moral judgments were innate but *subjective* the same form of argument might apply.[4] But this evolutionary debunking argument does not go so far as to endorse thesis (2), thus leaving open the possibility that justification may be reinstated. This epistemological undermining is in a sense provisional: it represents a *challenge* that the supporter of moral belief must rise to in order to overcome. One way of meeting the challenge would be to defend a form of moral naturalism according to which moral facts are identical to or supervene upon the items accepted in the evolutionary genealogy of morals. Therefore, on an earlier occasion, I supplemented my challenge with an attack on moral naturalism, centered on the contention that morality is essentially imbued with a kind of practical authority (conceptually speaking) which no combination of natural properties can provide. Thus while I have argued that empirical evolutionary discoveries are sufficient to create a substantive burden for the moralist, I also recognize the need to appeal to a priori metaethical methods in bolstering the challenge.

There are many additional nuanced ways that evolutionary findings might influence moral philosophy. The foregoing discussion of this section has confined itself to outlining two very broad programs – vindication and debunking – that purport to draw weighty ethical conclusions from biological discoveries, and identifying significant challenges for both.[5]

Notes

1 Following Ben Fraser (2010), I take spandrels to be traits that *necessarily* accompany adaptational mechanisms (like Stephen Jay Gould's original architectural spandrels, which are *unavoidable* features of a cathedral built with such-and-such a functional design); whereas *by-products* are traits that the adaptational mechanisms make possible but not inevitable. For the original metaphor, see Gould and Lewontin (1979).

2 I introduce the term "exaptation" with some hesitancy, since a great deal of confusion surrounds it (see Buss *et al.* 1998 for excellent analysis). Biological natural selection, as we have seen, can produce both adaptations and spandrels/by-products. Both kinds of trait can be co-opted for new uses, and when they are they then become *exaptations*, to use the term introduced by Gould (Gould and Vrba 1982; Gould 1991). Gould is unclear about what processes can perform this "co-opting," but it seems reasonable to maintain that when the process is natural selection then the exaptation is also an adaptation. Feathers, for example, possibly evolved originally as devices for thermoregulation; this was their adaptive function. Gradually, however, in certain organisms they were co-opted for a new function: they aided gliding and ultimately flight. While this permits one to describe modern avian feathers as an "exaptation," it seems unnecessarily counterintuitive to assume that it thereby excludes referring to them as an "adaptation." Dennett has complained that really *all* adaptations are exaptations (1995: 281), but

David Buss and colleagues have plausibly argued that there is some explanatory benefit in maintaining a distinction (see Buss *et al.* 1998: 542). However, there is no need to see exaptation and adaptation as contrary categories; the distinction can be maintained perfectly well if one classes the former as a proper subset of the latter. See Fraser (2010) who usefully discusses the place of exaptation in the moral nativist debate.

3 On other occasions I have certainly accepted something close to thesis (4) (see Joyce 2001), but did not attempt to establish it on evolutionary grounds.

4 For example, suppose we found ourselves believing that some actions are required because they are commanded by an all-powerful divine entity. The *nonobjectivity* of these requirements (something which can be maintained because the requirements depend on some entity's mental states) would not prevent the deployment of a genealogical debunking argument. For example, any plausible adaptational hypothesis concerning why our ancestors profited from these beliefs would presumably not make reference to their successfully tracking the mental states of any such divine entity. Confirmation of such a hypothesis would thus account for our beliefs in a way that neither implies nor presupposes their truth; and this, it can be argued, has implications for the epistemological status of these beliefs.

5 Thanks both to Ben Fraser and the editors of this collection for useful comments.

References

Alexander, R. (1987) *The Biology of Moral Systems*, Hawthorne, NY: Aldine de Gruyter.

Ayala, F. (2010) "The Difference of Being Human: Morality," *Proceedings of the National Academy of Sciences* 107: 9015–22.

Boyd, R. and Richerson, P.J. (1985) *Culture and the Evolutionary Process*, University of Chicago Press.

Brown, S. (2008) *Moral Virtue and Nature*, New York: Continuum.

Buller, D. (2006) *Adapting Minds*, Cambridge, MA: MIT Press.

Buss, D., Haselton, M., Shackleford, T., Bleske, A., and Wakefield, J. (1998) "Adaptations, Exaptations, and Spandrels," *American Psychologist* 53: 533–48.

Campbell, R. (1996) "Can Biology Make Ethics Objective?" *Biology and Philosophy* 11: 21–31.

Casebeer, W.D. (2003) *Natural Ethical Facts*, Cambridge, MA: MIT Press.

Chomsky, N. (1967) "Recent Contributions to the Theory of Innate Ideas," *Synthese* 17: 2–11.

Chomsky, N. (1987/1990) "On the Nature, Use, and Acquisition of Language," in *Mind and Cognition*, ed. W. Lycan, Oxford: Blackwell, pp. 627–46.

Confer, J., Easton, J., Fleischman, D., Goetz, C., Lewis, D., Perilloux, C., and Buss, D. (2010) "Evolutionary Psychology: Controversies, Questions, Prospects, and Limitations," *American Psychologist* 65: 110–26.

Conway, L. III and Schaller, M. (2002) "On the Verifiability of Evolutionary Psychological Theories: An Analysis of the Psychology of Scientific Persuasion," *Personality and Social Psychology Review* 6: 152–66.

Darwin, C. (1879/2004) *The Descent of Man*, London: Penguin Books.

Dennett, D.C. (1995) *Darwin's Dangerous Idea*, New York: Simon & Schuster.
de Waal, F.B.M. (1992) "The Chimpanzee's Sense of Social Regularity and Its Relation to the Human Sense of Justice," in *The Sense of Justice: Biological Foundations of Law*, eds. R.D. Masters and M. Gruter, Newbury Park, CA: Sage Publications, pp. 241–55.
de Waal, F.B.M. (1996) *Good Natured: The Origins of Right and Wrong in Primates and Other Animals*, Cambridge, MA: Harvard University Press.
de Waal, F.B.M. (2006) *Primates and Philosophers*, Princeton University Press.
Dwyer, S. (2006) "How Good is the Linguistic Analogy?" in *The Innate Mind, Volume 2: Culture and Cognition*, eds. P. Carruthers, S. Laurence, and S. Stich, Oxford: Oxford University Press, pp. 237–55.
Foot, P. (1972) "Morality as a System of Hypothetical Imperatives," *Philosophical Review* 81: 305–16.
Frank, R.H. (1988) *Passions within Reason: The Strategic Role of the Emotions*, New York: W.W. Norton & Company.
Fraser, B. (2010) "Adaptation, Exaptation, By-products and Spandrels in Evolutionary Explanations of Morality," *Biological Theory* 5: 223–7.
Garner, R. (2010) "Abolishing Morality," in *A World without Values*, eds. R. Joyce and S. Kirchin, Dordrecht: Springer Press, pp. 217–33.
Gould, S.J. (1991) "Exaptation: A Crucial Tool for Evolutionary Psychology," *Journal of Social Issues* 47: 43–65.
Gould, S.J., ed. (1992) "Male Nipples and Clitoral Ripples," in *Bully for Brontosaurus: Further Reflections in Natural History*, New York: W.W. Norton & Company, pp. 124–38.
Gould, S.J. and Lewontin, R.C. (1979) "The Spandrels of San Marco and the Panglossion Paradigm: A Critique of the Adaptationist Programme," *Proceedings of the Royal Society: Biological Sciences, Series B* 205: 581–98.
Gould, S.J. and Vrba, E. (1982) "Exaptation: A Missing Term in the Science of Form," *Paleobiology* 8: 4–15.
Greene, J. (2003) "From Neural 'Is' to Moral 'Ought': What are the Moral Implications of Neuroscientific Moral Psychology?" *Nature Reviews Neuroscience* 4: 847–50.
Griffiths, P. (2002) "What is Innateness?" *The Monist* 85: 70–85.
Haidt, J. and Bjorkland, F. (2008) "Intuitive Ethics: How Innately Prepared Intuitions Generate Culturally Variable Virtues," in *Moral Psychology, Volume 2: The Cognitive Science of Morality: Intuition and Diversity*, ed. W. Sinnott-Armstrong, Cambridge, MA: MIT Press, pp. 181–217.
Haidt, J. and Joseph, C. (2004) "Intuitive Ethics: How Innately Prepared Intuitions Generate Culturally Variable Virtues," *Daedalus* 133: 55–66.
Hamlin, J.K., Wynn, K., and Bloom, P. (2007) "Social Evaluation by Preverbal Infants," *Nature* 450: 557–9.
Hamlin, J.K., Wynn, K., and Bloom, P. (2010) "Three-Month-Old Infants Show a Negativity Bias in Social Evaluation," *Developmental Science* 13: 923–9.
Hauser, M. (2006) *Moral Minds*, New York: Harper Collins.
Hinckfuss, I. (1987) "The Moral Society: Its Structure and Effects," *Discussion Papers in Environmental Philosophy* 16. Canberra: Philosophy Program (RSSS), Australian National University.
Joyce, R. (2001) *The Myth of Morality*, Cambridge: Cambridge University Press.
Joyce, R. (2006) *The Evolution of Morality*, Cambridge, MA: MIT Press.

Joyce, R. (2013) "The Many Moral Nativisms," in *Cooperation and its Evolution*, eds. K. Sterelny, R. Joyce, B. Calcott, and B. Fraser, Cambridge, MA: MIT Press.

Kahane, G. (2011) "Evolutionary Debunking Arguments," *Noûs* 45: 103–25.

Kant, I. (1783/1985) *Groundwork to the Metaphysics of Morals*, trans. H.J. Paton, London: Hutchinson.

Ketelaar, T. and Ellis, B. (2000) "Are Evolutionary Explanations Unfalsifiable? Evolutionary Psychology and the Lakatosian Philosophy of Science," *Psychological Inquiry* 11: 1–21.

Kitcher, P. (2011) *The Ethical Project*, Cambridge, MA: Harvard University Press.

Krebs, D. (2005) "The Evolution of Morality," in *The Handbook of Evolutionary Psychology*, ed. D. Buss, Hoboken, NJ: John Wiley & Sons, Inc., pp. 747–71.

Laurence, S. and Margolis, E. (2001) "The Poverty of the Stimulus Argument," *British Journal for the Philosophy of Science* 52: 217–76.

Mackie, J.L. (1977) *Ethics: Inventing Right and Wrong*, London: Penguin.

Mameli, M. and Bateson, P. (2007) "The Innate and the Acquired: Useful Clusters or a Residual Distinction from Folk Biology?" *Developmental Psychobiology* 49: 818–31.

Mason, K. (2010) "Debunking Arguments and the Genealogy of Religion and Morality," *Philosophy Compass* 5: 770–8.

Mikhail, John (2011) *Elements of Moral Cognition: Rawls' Linguistic Analogy and the Cognitive Science of Moral and Legal Judgment*, Cambridge: Cambridge University Press.

Miller, G. (2000) *The Mating Mind*, Doubleday.

Miller, G. (2007) "Sexual Selection for Moral Virtues," *Quarterly Review of Biology* 82: 97–121.

Nesse, R. (2007) "Runaway Social Selection for Displays of Partner Value and Altruism," *Biological Theory* 2: 143–55.

Nichols, S. (2005) "Innateness and Moral Psychology," in *The Innate Mind: Structure and Contents*, eds. P. Carruthers, S. Laurence, and S. Stich, New York: Oxford University Press, pp. 353–430.

Nietzsche, F. (1887/1994) *The Genealogy of Morals*, trans. C. Diethe and ed. K. Ansell-Pearson, Cambridge: Cambridge University Press.

Noë, R. (2001) "Biological Markets: Partner Choice as the Driving Force behind the Evolution of Mutualisms," in *Economics in Nature: Social Dilemmas, Mate Choice, and Biological Markets*, eds. R. Noë, J. van Hoof, and P. Hammerstein, Cambridge: Cambridge University Press, pp. 93–118.

Nucci, L.P. (2001) *Education in the Moral Domain*, Cambridge: Cambridge University Press.

Prinz, J. (2008) "Is Morality Innate?" in *Moral Psychology, Volume 1: The Evolution of Morality: Adaptations and Innateness*, ed. W. Sinnott-Armstrong, Cambridge, MA: MIT Press, pp. 367–406.

Prinz, J. (2009) "Against Moral Nativism," in *Stich and His Critics*, eds. D. Murphy and M. Bishop, Oxford: Wiley-Blackwell, pp. 167–89.

Richards, R.J. (1986) "A Defense of Evolutionary Ethics," *Biology and Philosophy* 1: 265–93.

Richerson, P.J. and Boyd, R. (2005) *Not by Genes Alone: How Culture Transformed Human Evolution*, Chicago: University of Chicago Press.

Rottschaefer, W.A. and Martinsen, D. (1990) "Really Taking Darwin Seriously: An Alternative to Michael Ruse's Darwinian Metaethics," *Biology and Philosophy* 5: 149–73.

Ruse, M. (1986) *Taking Darwin Seriously*, Oxford: Basil Blackwell.
Ruse, M. (2006) "Is Darwinian Metaethics Possible (and if It is, is It Well-Taken)?" in *Evolutionary Ethics and Contemporary Biology*, eds. G. Boniolo and G. de Anna, Cambridge: Cambridge University Press, pp. 13–26.
Ruse, M., ed. (2009) "Evolution and Ethics: The Sociobiological Approach," in *Philosophy after Darwin*, Princeton, NJ: Princeton University Press, pp. 489–511.
Ruse, M. and Wilson, E.O. (1985) "The Evolution of Ethics," *New Scientist* 108: 50–2.
Singer, P. (2005) "Ethics and Intuitions," *Journal of Ethics* 9: 331–52.
Smetana, J.G. (1993) "Understanding of Social Rule," in *The Development of Social Cognition: The Child as Psychologist*, ed. M. Bennett, New York: Guildford Press, pp. 111–41.
Smetana, J.G. and Braeges, J.L. (1990) "The Development of Toddlers' Moral and Conventional Judgments," *Merrill-Palmer Quarterly* 36: 329–46.
Sober, E. (2000) "Psychological Egoism," in *The Blackwell Guide to Ethical Theory*, 1st edn, ed. H. LaFollette, Oxford: Blackwell, pp. 129–48.
Sober, E. and Wilson, D.S. (1998) *Unto Others: The Evolution and Psychology of Unselfish Behavior*, Cambridge, MA: Harvard University Press.
Spencer, H. (1851) *Social Statics*, London: J. Chapman.
Spencer, H. (1879/2011) *The Data of Ethics*, ed. J. Turner, New Jersey: Transaction.
Sterelny, K. (2010) "Moral Nativism: A Sceptical Response," *The Sense of Justice: Biological Foundations of Law* 25: 279–97.
Stich, S. (2007) "Evolution, Altruism and Cognitive Architecture: A Critique of Sober and Wilson's Argument for Psychological Altruism," *Biology and Philosophy* 22: 267–81.
Street, S. (2006) "A Darwinian Dilemma for Realist Theories of Value," *Philosophical Studies* 127: 109–66.
Tomkins J.L. and Brown, G.S. (2004) "Population Density Drives the Local Evolution of a Threshold Dimorphism," *Nature* 431: 1099–103.
Turiel, E. (2002) *The Culture of Morality: Social Development, Context, and Conflict*, Cambridge: Cambridge University Press.
Weinstein, D. (1998) *Equal Freedom and Utility: Herbert Spencer's Liberal Utilitarianism*, Cambridge: Cambridge University Press.
Westermarck, E. (1906) *The Origin and Development of the Moral Ideas*, vol. 2, London: Macmillan.
Williams, G.C. (1966) *Adaptation and Natural Selection*, Princeton University Press.
Wilson, E.O. (1978) *On Human Nature*, Cambridge, MA: Harvard University Press.
Zahavi, A. (1977) "The Cost of Honesty: Further Remarks on the Handicap Principle," *Journal of Theoretical Biology* 67: 603–5.

Chapter 7

Psychological Egoism

Elliott Sober

Psychological egoism is a theory about motivation. It claims that all of our ultimate desires are self-directed. Whenever we want others to do well (or ill), we have these other-directed desires only instrumentally; we care about others only because we think that the welfare of others will have ramifications for our own welfare. As stated, egoism is a descriptive, not a normative, claim. It aims to characterize what motivates human beings in fact; the theory does not say whether it is good or bad that people are so motivated.

Egoism has exerted a powerful influence in the social sciences and has made large inroads in the thinking of ordinary people. Economists often think of human beings as being moved by "rational self-interest," where this excludes any irreducible concern for the welfare of others. And ordinary folks often claim that people help others only because this makes them feel good about themselves, or because they seek the approval of third parties.

It is easy to invent egoistic explanations for even the most harrowing acts of self-sacrifice. The soldier in a foxhole who throws himself on a grenade to save the lives of his comrades is a fixture in the literature on egoism. How could this act be a product of self-interest if the soldier knows that it will end his life? The egoist may answer that the soldier realizes in an instant that he would rather die than suffer the guilt feelings that would haunt him if he saved himself and allowed his friends to perish. The soldier prefers to die and then have no sensations at all rather than live and suffer the torments of the damned. This reply may sound forced, but it remains to be seen what grounds we have for regarding it as false.

The criticisms that have been leveled against psychological egoism can be divided into three categories. First, there is the claim that it is not a genuine theory at all. Second, there is the allegation that it is a theory that is refuted by what we

The Blackwell Guide to Ethical Theory, Second Edition. Edited by Hugh LaFollette and Ingmar Persson.

observe in human behavior. Third, there is the idea that, although egoism is a theory that is consistent with what we observe, there are other, extra-evidential considerations that suggest that it should be rejected in favor of an alternative theory, motivational pluralism, according to which human beings have both egoistic and altruistic ultimate desires. All three types of criticism will be considered in what follows, but first we need to state the theory more carefully.

Clarifying Egoism

When egoism claims that all our ultimate desires are self-directed, what do "ultimate" and "self-directed" mean?

There are some things that we want for their own sakes; other things we want only because we think they will get us something else. The familiar means/end relation that links one desire to another also allows desires to be chained together – Sarah may want to drive her car because she wants to get to the bakery; she may want to go to the bakery because she wants to buy bread; and so on. The crucial relation that we need to define is this:

> *S* wants *m* solely as a means to acquiring *e* if and only if *S* wants *m*, *S* wants *e*, and *S* wants *m* only because she believes that obtaining *m* will help her obtain *e*.

An ultimate desire is simply a desire that someone has for reasons that go beyond its ability to contribute instrumentally to the attainment of something else. Consider pain. The most obvious reason that people want to avoid pain is simply that they dislike experiencing it. Avoiding pain is one of our ultimate goals. However, many people realize that being in pain reduces their ability to concentrate, so they may sometimes take an aspirin in part because they want to remove a source of distraction. This shows that the things we want as ends in themselves we also may want for instrumental reasons.

When psychological egoism seeks to explain why one person helped another, it is not enough to show that *one* of the reasons for helping was self-benefit; this is quite consistent with there being another, purely altruistic, reason that the individual had for helping. Symmetrically, to refute egoism, one need not cite examples of helping in which only other-directed motives play a role. If people sometimes help for both egoistic and altruistic ultimate reasons, then psychological egoism is false.

Egoism and altruism both require the distinction between self-directed and other-directed desires. This distinction is to be understood in terms of a desire's propositional content. If Adam wants the apple, this is elliptical for saying that Adam wants it to be the case that *he has the apple*. This desire is purely self-directed, since its propositional content mentions Adam, but no other agent; I assume that Adam does not regard the apple as an agent. In contrast, when Eve wants *Adam*

to have the apple, this desire is purely other-directed; its propositional content mentions another person, Adam, but not Eve herself. Egoism claims that all of our ultimate desires are self-directed; altruism, that some are other-directed. The fact that Eve has an other-directed desire is not enough to refute egoism; one must ask *why* Eve wants Adam to have the apple.

A special version of egoism is psychological hedonism. The hedonist says that the only ultimate desires that people have are attaining pleasure and avoiding pain. Hedonism is sometimes criticized for holding that pleasure is a single type of sensation – that the pleasure we get from the taste of a peach and the pleasure we get from seeing the prospering of those we love somehow boil down to the same thing (LaFollette 1988). However, this criticism does not apply to hedonism as I have described it. The important point about this theory is its claim that people are motivational solipsists; the only things they care about ultimately are states of their own consciousness. Egoists need not be hedonists. If people desire their own survival as an end in itself, they may be egoists, but they are not hedonists.

There are desires that are neither purely self-directed nor purely other-directed. If Phyllis wants to be famous, this means that she wants others to know who she is. This desire's propositional content involves a relation between self and others. If Phyllis seeks fame solely because she thinks this will bring her pleasure or profit, then she may be an egoist (depending on what her other ultimate desires happen to be). But what if she wants to be famous as an end in itself? There is no reason to cram this possibility within egoism or altruism; to include some ultimate relational desires, but not others, within egoism, runs the risk of making the theory ad hoc or unclear (Kavka 1986); and the same point also applies to altruism. So let us recognize *relationism* as a possibility distinct from both.

A fourth possibility involves desires that mention neither self nor other. The desire that some general moral principle be upheld falls into this category. When a utilitarian desires the greatest good for the greatest number, the desire is impersonal; the desire covers all sentient beings, presumably including the desirer himself, but the desire's content singles out neither self nor specific others. For this reason, I suggest that it is neither altruistic nor egoistic. Just as was true with respect to relational desires, the defender of psychological egoism can grant that there are desires concerning general moral principles that are not self-directed; the question is whether we have these desires instrumentally or as ends in themselves.

With egoism characterized in the way I have suggested, it obviously is not entailed by the truism that people act on the basis of their own desires, nor by the truism that they seek to have their desires satisfied. The fact that Joe acts on the basis of Joe's desires, not on the basis of Jim's, tells us whose desires are doing the work; it says nothing about whether the ultimate desires in Joe's head are purely self-directed. And the fact that Joe wants his desires to be satisfied means merely that he wants their propositional contents to come true; Joe's desire that it rain tomorrow is satisfied if and only if it rains tomorrow (Stampe 1994). If

there is rain, the desire is satisfied, whether or not Joe knows that it is. To want one's desires satisfied is not the same as wanting the feeling of satisfaction that sometimes accompanies a satisfied desire.

Egoism is sometimes criticized for attributing too much calculation to spontaneous acts of helping. People who help in emergency situations often report doing so "without thinking" (Clark and Word 1974). However, it is hard to take such reports literally when the acts involve a precise series of complicated actions that are well suited to an apparent end. A lifeguard who rescues a struggling swimmer is properly viewed as having a goal and as selecting actions that advance that goal. The fact that she engaged in no ponderous and self-conscious calculation does not show that no means/end reasoning occurred. In any case, actions that really do occur without the mediation of beliefs and desires fall outside the scope of both egoism and altruism. People jerk their legs when their knees are tapped with hammers, but that refutes neither theory.

A related criticism is that egoism assumes that people are more rational than they really are. However, recall that egoism is simply a claim about the ultimate desires that people have. As such, it says nothing about how people decide what to do on the basis of their beliefs and desires. Theorists who assume that egoism is true also often assume that people are rational calculators; however, theories are not convicted by a principle of guilt by association. The assumption of rationality is no more a part of psychological egoism than it is part of motivational pluralism.

If egoism holds that all ultimate desires are self-directed, what are we to say of someone whose ultimate goal is his own destruction? And if altruism holds that some of our ultimate desires are other-directed, what are we to make of Iago, who has the ultimate goal of destroying Othello? It is jarring to say that a depressed person bent on suicide is an egoist, or that Iago is an altruist. What we need to add to both theories is the idea of what is good (or apparently good). Egoists seek their own benefit; altruists want others to do well. Although these additions to the theories bring them more in line with ordinary usage of the terms "egoism" and "altruism," they do not materially affect the substantive task of determining which theory is true. The crux of the problem is to tell whether all ultimate desires are self-directed.

It may strike some readers that the problem is easy. Individuals can merely gaze within their own minds and determine by introspection what their ultimate motives are. Perhaps advocates of egoism are right about themselves and advocates of motivational pluralism are right about themselves; both sides err only when they generalize beyond their own cases. An implicit assumption, in both philosophical and psychological explorations of this topic, is that people are basically the same. If egoism is false, it is false for practically everyone (sociopaths, perhaps, excepted). And if it is true, it is true because it characterizes a basic feature of human nature. However, the fact that earlier work in psychology and philosophy often ignored the possibility of individual variation is no reason to build this into our understanding of the problem. Why, then, should we not say that advocates of egoism know

their own hearts and that defenders of altruism know theirs? The reason is that there is no independent reason to think that the testimony of introspection is to be trusted in this instance. Introspection is misleading or incomplete in what it tells us about other facets of the mind; no one has shown why the mind must be an open book with respect to this question about ultimate motives. The problem, if it can be solved, must be solved in some other way.

Is Egoism Empirically Testable?

One standard philosophical objection to egoism is that it is not a testable hypothesis. As the example of the soldier in the foxhole suggests, egoism seems able to accommodate any behavior whatever. Whether people are nasty or nice to each other, the theory can explain why. This claim about the flexibility of egoism is then linked to a Popperian criterion concerning what it takes for a statement to be scientific, with the conclusion drawn that egoism is not a genuine scientific theory at all. It is, despite appearances, empirically vacuous.

This argument is flawed in two ways. The first pertains to its sanguine confidence that no observation could ever disconfirm egoism. The fact that the theory can accommodate the soldier in the foxhole and other behaviors that have been considered by philosophers hardly suffices to justify this global claim. As it happens, experimental work in social psychology on altruism and egoism shows that the relevant observational evidence extends beyond the existence of instances of helping behavior (Batson 1991; Schroeder *et al.* 1995). In addition, the Duhemian point that theories are testable only in conjunction with background assumptions should lead us to draw back from this charge of untestability. If two theories make the same predictions in conjunction with one set of background assumptions, they may make different predictions when conjoined to another. How do we know that new background theories will never be developed that allow egoism to be put to the test? The charge of untestability presupposes that we have an omniscient grasp of the future of science.

The second defect in this argument is that it neglects to notice that the charge of untestability is a two-edged sword. The argument is advanced as a reason for rejecting egoism. What, then, are we to accept as a positive account of motivation? Presumably, motivational pluralism is supposed to be the acceptable alternative. However, this cannot be where the argument leads. If egoism is untestable, then so is motivational pluralism. As flexible as egoism is in its ability to accommodate observations, pluralism is more flexible still. After all, pluralism deploys all the variables that egoism invokes, and then some.

The reason egoism appears to be untestable is that it is an *ism*. It does not provide specific explanations for behaviors, but merely indicates the kind of explanation that all behaviors will have. This is why it is possible for egoism to be

retained even when specific egoistic explanations are found wanting. Why did George donate all that money to charity? A defender of egoism might suggest that George did so because he wanted to improve his business contacts by impressing others. But suppose one then learns that George donated the money anonymously. This refutes the specific egoistic explanation just described, but it is not hard to invent another. George made the donation because it made him feel good and because he knew that if he did not, he would experience pangs of guilt. The pattern here is typical – hedonism is the position to which egoists standardly retreat. If external benefits do not suffice to explain, one invokes internal, psychological benefits instead.

That egoism is a claim about a *type* of explanation, and therefore is distinct from the *specific explanations* that are of the type required, is a pattern that arises in many debates about isms. Consider adaptationism in evolutionary biology. Adaptationists emphasize the importance of natural selection in explaining the observed traits of organisms. Because this ism, by itself, does not provide a specific explanation for any trait, it remains possible for a biologist to continue to be an adaptationist even after a specific adaptationist explanation is found wanting. Why did wings evolve in insects? The hypothesis that wings evolved as an adaptation for flying is thrown in doubt by the fact that very small wing buds provide no lift whatever; although 5% of an eye can still function as a light sensor, 5% of a wing does nothing to get an organism off the ground. But it turns out that wing buds are found in some flightless insect species; the buds function as thermoregulators. This suggests an alternative adaptationist hypothesis – that insect wings started to evolve because they initially promoted thermoregulation and then continued to evolve because they then facilitated flight. And if this hypothesis is challenged, the adaptationist can cast about for a third alternative. It is no good rejecting adaptationism because it has this sort of flexibility; the alternative ism, evolutionary pluralism, claims that natural selection is one among several important causes of evolution. As flexible as adaptationism is, pluralism is more flexible still.

Butler's Stone

As noted in the previous section, even though hedonism is a special version of egoism, hedonistic explanations are often what egoists invoke when a nonhedonistic explanation is found wanting. If George did not donate money to charity to make business contacts, perhaps he did so for the warm glow of satisfaction that the donation provided. For this reason, arguments that attempt to refute hedonism have a special location in the dialectical landscape. Although refuting hedonism is not sufficient to refute egoism, it would make an important contribution to that larger enterprise.

Many philosophers have thought that Joseph Butler (1692–1752) refuted hedonism once and for all (Broad 1965; Feinberg 1984; Nagel 1970) in the following passage:

> That all particular appetites and passions are towards *external things themselves*, distinct from the *pleasure arising from them*, is manifested from hence; that there could not be this pleasure, were it not for that prior suitableness between the object and the passion: there could be no enjoyment or delight from one thing more than another, from eating food more than from swallowing a stone, if there were not an affection or appetite to one thing more than another.
>
> (Butler 1726/1965: 227)

I will call this argument *Butler's stone*. Although Butler does not explicitly say in this passage that hedonism is false, let us construe the argument with this as its conclusion:

(1) People sometimes experience pleasure.
(2) When people experience pleasure, this is because they had a desire for some external thing, and that desire was satisfied.

___

Hedonism is false.

I do not propose to challenge the first premise. However, I think the second premise is false and that the conclusion does not follow from the premises.

The second premise is overstated; although some pleasures are the result of a desire's being satisfied, others are not (Broad 1965: 66). One can enjoy the smell of violets without having formed the desire to smell a flower or something sweet. Since desires are propositional attitudes, forming a desire is a cognitive achievement. Pleasure and pain, on the other hand, are sometimes cognitively mediated, but sometimes they are not. Notice that this defect in the argument can be repaired: Butler does not need to say that desire satisfaction is the one and only road to pleasure.

The transition from premises to conclusion is where the argument really goes wrong. Consider the causal chain from a *desire* (the desire for food, say), to an *action* (eating), to a *result* – pleasure. Because the pleasure traces back to an antecedently existing desire, it will be false that the resulting pleasure caused the desire (on the assumption that cause precedes effect). However, this does not settle how two *desires* – the *desire for food* and the *desire for pleasure* – are related. In particular, it leaves entirely open what caused the desire for food. Hedonism says that people desire food *because* they want pleasure (and think that food will bring them pleasure). Butler's stone concludes that this causal claim is false, but for no good reason. The crucial mistake in the argument comes from confusing two quite different items – the *pleasure* that results from a desire's being satisfied and the *desire for pleasure*. Even if the occurrence of pleasure presupposed that the agent desired

something besides pleasure, nothing follows about the relationship between the *desire for pleasure* and the desire for something else (Sober 1992; Stewart 1992; Sober and Wilson 1998). Hedonism does not deny that people desire external things; rather, the theory tries to explain why that is so.

It is curious that this argument has been interpreted so widely as refuting hedonism. At the end of the sermon in which the stone passage occurs, Butler says this:

> Let it be allowed, though virtue or moral rectitude does indeed consist in affection to and pursuit of what is right and good, as such; yet, that when we sit down in a cool hour, we can neither justify to ourselves this or any other pursuit, till we are convinced that it will be for our happiness, or at least not contrary to it.
>
> (Butler 1726/1965: 240)

And if we return to the language of the stone argument itself, we see that Butler is making a claim about the content of "particular appetites and passions." Read narrowly, the argument says merely that if people desire pleasure, their desires do not fall under that rubric. The argument does not say that people never desire pleasure; nor does it say that the desire for pleasure is never ultimate. Did Butler fail to refute hedonism in the stone argument because he never tried to do so?

The "Paradox" of Hedonism and Its "Irrationality"

Individuals who focus exclusively on attaining pleasure or happiness inevitably fail to get what they want. They are like stockbrokers who think only that they should buy low and sell high. People who have an end in view but never consider what means they should use to pursue their goal surely will fail to get what they want. This has led some philosophers to claim that pleasure and happiness are attainable only as by-products of becoming absorbed in specific activities. They also have suggested that this fact about pleasure and happiness constitutes a paradox for hedonism – the word "paradox" indicating that we are supposed to find here a flaw in hedonism as a psychological theory (Butler 1726/1965; Feinberg 1984).

The obvious reply to this criticism is that there is nothing in hedonism that says that people must be monomaniacs. Hedonism says that people have attaining pleasure and avoiding pain as their only *ultimate* goals; it does not say that attaining pleasure and avoiding pain are the only goals (ultimate *or* proximate) that people ever have. Hedonists reflect on which activities are most apt to bring pleasure and prevent pain, and decide what to do on that basis (Sidgwick 1907/1922). Furthermore, if hedonistic monomaniacs always fail to get what they want, what follows from this? Even if this entailed that people *should* not be hedonists, it does not show that people are not hedonists *in fact* – recall that hedonism is a descriptive, not a normative, theory.

The normative/descriptive distinction is also relevant to evaluating the claim that egoism is irrational. Nagel (1970) defends this claim by contending that when egoists consider their own interests in deliberation, but not those of others, they neglect the fact that there is no property that they have and others lack that could justify this asymmetry. To evaluate whether egoists are irrational, we need to decide whether rationality should be understood "instrumentally" or "substantively." Instrumental rationality just means the ability to choose efficient means to achieve whatever ends one might have. The substantive notion means not just that efficient means have been secured but that the ends are praiseworthy, or at least are morally unobjectionable (Gibbard 1990). Efficient serial killers might be instrumentally rational, but they are not substantively rational. Regardless of which notion captures what the word "rational" means, the fact remains that this line of argument cannot show that people really have or are capable of having altruistic ultimate motives. If rationality just means instrumental rationality, then rationality does not entail altruism (or its possibility); if rationality means substantive rationality, then even if rationality entails altruism, it needs to be shown that people really are substantively rational. Perhaps we *ought* to be rational and maybe we *ought* to be altruistic as well. This does not show that egoism is false as a descriptive thesis.

The Experience Machine

In the science fiction movie *Total Recall*, people centuries from now use their computer technology to go on "virtual vacations." Instead of going on a real vacation, they plug into a computer that provides a thoroughly convincing simulation of a real vacation. The movie quite plausibly suggests that people in the future often might choose to "vacation" in this way, especially if real trips to exotic locales are expensive and dangerous, while "virtual vacations" are cheap and completely convincing from an experiential point of view.

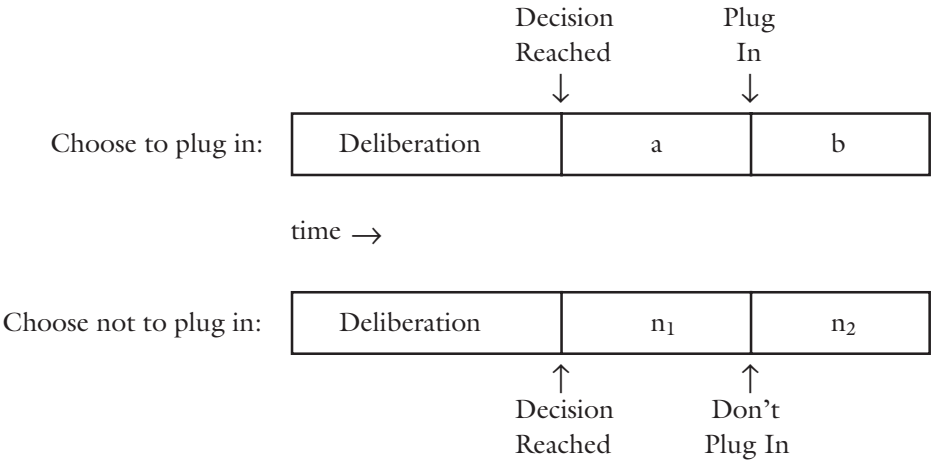
Robert Nozick wrote *Anarchy, State, and Utopia* before *Total Recall* appeared. In his book he uses the idea of an "experience machine" to construct an argument that seems to show that hedonism is false (Nozick 1974: 42–5). Nozick's machine can be programmed to provide thoroughly convincing simulations of any real-life experience one might choose. Suppose you were offered the chance to plug into the experience machine for the rest of your life. The machine would be programmed to make you instantly forget that you had chosen to plug in and then the machine would provide whatever sequence of experiences you would find maximally pleasurable and minimally painful. Of course, your beliefs about the type of life you are leading will be false. If you choose to plug into the experience machine, you will live your life strapped to a laboratory table with tubes and electrodes sticking into your body. You will never *do* anything; however, the level of pleasure you will experience, thanks to the machine, will be extraordinary.

If you were offered the chance to plug into the experience machine for the rest of your life, what would you do? Your first reaction might be to doubt that the machine will perform as promised; certainly no machine now on the market can deliver what this machine is said to be able to do, and this will remain true at least for the foreseeable future. However, for the sake of argument, try to set this hesitation to one side. Imagine yourself being offered the chance to plug in, and suppose that the machine will work as described. My guess is that many people, perhaps yourself included, would decline the opportunity of plugging in.

This fact about people seems to refute hedonism. Apparently, many people prefer to have a real life over a simulated one, even if real life brings less pleasure and more pain than the life they would have if they plugged into the machine. It seems that people care irreducibly about how they are related to the world outside their own minds; it is false that the only things they care about as ends in themselves are pleasant states of consciousness.

Can hedonism explain why many people would decline the offer to plug into the machine? To see whether this is possible, we need to map out the sequence of events that will comprise your life if you choose to plug into the experience machine and the sequence of events that will occur if you do not. In both cases, the process begins with deliberation, which terminates in a decision. If you decide to plug in, there is a time lag between your decision and you actually being connected to the machine. Figure 7.1 shows the two time lines we need to consider. The four letters in these two time lines represent how pleasant your experiences will be during different temporal periods, depending on what you decide. If you choose to plug into the machine, you will have an immense level of *bliss* (*b*) after you plug in. This will dwarf the amount of pleasure you will

Figure 7.1



experience in the same period of time if you decide not to plug in and to lead a normal life instead: $b > n_2$. If this were the only consideration involved, the hedonist would have to predict that people will choose to plug into the machine. How can hedonism explain the fact that many people make the opposite decision?

The hedonist's strategy is to look at earlier events. If you decided to plug into the machine, how would you feel before you were actually connected? Presumably, you would experience a great deal of *anxiety* (a). You would realize that you were about to stop leading a real life. You will never again see the people you love; all of your projects and plans are about to be terminated. It is clear that you would have less pleasure during this period of time than you would if you rejected the option of plugging into the machine and continued with your real life instead: $a < n_1$.

If hedonists are to explain why people choose not to plug into the experience machine, and are to do this by considering just the pleasure and pain that subjects expect to come their way *after* they decide what to do, the claim must be that $a + b < n_1 + n_2$. Since b is far greater than n_2, this inequality will be true only if a is far far smaller than n_1. That is, hedonists seem compelled to argue that people reject the option of plugging in because the amount of pain they would experience between deciding to plug in and actually being connected to the machine is *gigantic* – so large that it dwarfs the pleasure they would experience after they are connected.

This suggestion is not plausible. The period of time between deciding to plug in and actually doing so can be made very brief, compared with the long stretch of years you will spend attached to the machine and enjoying a maximally pleasurable ensemble of experiences. I grant that people who decide to connect to the machine will experience sadness and anxiety during the brief interval between deciding to plug in and actually plugging in. But the idea that this negative experience swamps all subsequent pleasures is just not credible.

To see why, let us consider a second thought experiment, suggested to me by William Talbott. Suppose you were offered $5,000 if you went through ten seconds of a certain experience. The experience is believing that you had just decided to spend the rest of your life plugged into an experience machine. After your ten second jolt of this experience, you will return to your normal life and will realize that you had just had a "nightmare"; you then will receive the money as promised. I expect that many people would choose the ten seconds just described because it will earn them $5,000. This shows that hedonism is mistaken if it claims that the experience of believing for a few minutes that you will be plugged into an experience machine for the rest of your life is so horrible that no one would ever choose a life that included it.

The hedonist still has not been able to explain why many people would choose a normal life over a life plugged into the experience machine. The reason is that a hedonistic calculation seems to lead inevitably to the conclusion that $a + b > n_1 + n_2$. Does this mean that the hedonist must concede defeat? I think that the hedonist has a way out. Quite apart from the amount of pleasure and pain

that accrues to subjects *after* they decide what to do, there is the level of pleasure and pain arising in the deliberation process itself. The hedonist can maintain that *deciding* to plug into the machine is so aversive that people almost always make the other choice. When people deliberate about the alternatives, they feel bad when they think of the life they will lead if they plug into the machine; they feel much better when they consider the life they will lead in the real world if they decline to plug in. The *idea* of life attached to the machine is painful, even though such a life would be quite pleasurable; the *idea* of real life is pleasurable, even though real life often includes pain. This hedonistic explanation of why people refuse to plug in exploits the distinction that Schlick (1939) drew between the pleasant idea of a state and the idea of a pleasant state.

To see what is involved in this suggestion, let us consider in more detail what goes through people's minds as they deliberate. They realize that plugging in will mean abandoning the projects and attachments they hold dear; plugging into the machine resembles suicide in terms of the utter separation it effects with the real world. The difference is that suicide means an end to consciousness, whereas the experience machine delivers (literally) escapist pleasures. Hedonism is not betraying its own principles when it claims that many people would feel great contempt for the idea of plugging in and would regard the temptation to do so as loathsome. People who decline the chance to plug in are repelled by the idea of narcissistic escape and find pleasure in the idea of choosing a real life.

One virtue of this hedonistic explanation is that it explains the results obtained in both the thought experiments described. It explains why people often *decline* to plug into the experience machine for the rest of their lives; it also explains why people offered $5,000 often *agree* to have ten seconds of the experience of believing that they have just decided to plug into the machine for the rest of their lives. In both cases deliberation is guided not so much by beliefs about which actions will bring *future* pleasure, but by the pleasure and pain that accompany certain thoughts *during the deliberation process itself*.

The problem of the experience machine resembles the problem of the soldier in the foxhole discussed earlier. How can hedonism explain this act of suicidal self-sacrifice if the soldier believes that he will not experience anything after he dies? The hedonist can suggest that there is a self-directed benefit that accrues *before* the act of self-sacrifice is performed. It is no violation of hedonism to maintain that the soldier decides to sacrifice his life because that decision is less painful than the decision to let his friends die. The problem of suicidal self-sacrifice and the problem posed by the experience machine can be addressed in the same way.

Burden of Proof

Philosophers sometimes maintain that a commonsense idea should be regarded as innocent until proven guilty. That is, if a question is raised about whether some

commonsense proposition is true, and no argument can be produced that justifies or refutes it, then the sensible thing to do is to keep on believing the proposition. Put differently, the idea is that the burden of proof lies with those who challenge common sense.

This general attitude sometimes surfaces in discussions of egoism and altruism. The claim is advanced that the egoism hypothesis goes contrary to common sense. The commonsense picture of human motivation is said to be pluralistic – people care about themselves, but they also care about others, not just as means, but as ends in themselves. The conclusion is then drawn that if philosophical and scientific argumentation for and against egoism is indecisive then we should reject egoism and continue to accept pluralism.

One objection to this proposed tiebreaker is that it is far from obvious that "common sense" is on the side of motivational pluralism rather than egoism. What is common sense? Isn't it just what people commonly believe? If so, it is arguable that egoism has made large inroads; it now seems to be a view that is endorsed by large numbers of people. Philosophers need to be careful not to confuse common sense with what they themselves happen to find obvious. As far as I know, no empirical survey has determined whether a pluralistic theory of motivation is more popular than psychological egoism.

Regardless of what people commonly believe about psychological egoism and motivational pluralism, I reject the idea that conformity with common sense is a tiebreaker in this debate. It does not have this status in physics or biology, and I see no reason why it should do so when the question happens to be philosophical or psychological in character. In fact, it is arguable that our intuitions in this domain are especially prone to error. If certain types of self-deception – either regarding one's own motives or those of others – were advantageous, then evolution might have enshrined these falsehoods in the set of "obvious" propositions we call common sense. A philosophy informed by an evolutionary perspective has no business taking common sense at face value.

Parsimony

I have so far argued that hedonism has not been refuted by philosophical arguments or by observed behavior; if this is right, then egoism has not been refuted either. This does not mean that egoism is true; after all, motivational pluralism has also not been refuted. In the light of this impasse, it is worth noting that social scientists often implicitly assume that if a behavior *can* be explained in egoistic terms, then it *ought* to be so explained. The fact that they have no observational argument in favor of egoism seems not to be relevant. And the fact that behavior also can be explained in terms of motivational pluralism also seems not to be relevant. This raises the question of why egoism is the default hypothesis – the hypothesis that we should assume is true unless we are forced to abandon it.

One answer to consider is that egoism is more parsimonious – it postulates only one type of ultimate motive, whereas pluralism postulates two (Hume 1751/1970; Batson 1991). Even if we assume that parsimony marks not just an aesthetic difference between theories, but a reason for finding some theories more plausible than others, there still is a defect in this defense of egoism. The problem is that egoism is *less* parsimonious than pluralism when we consider how many causal beliefs the two theories postulate. When Sally wants Otto to do well, the defender of egoism counts this as an instrumental desire while the proponent of motivational pluralism may hold that Sally has this other-directed desire as an end in itself. Notice that the egoistic explanation attributes to Sally a causal belief – *that she stands to receive a benefit from Otto's doing well.* Motivational pluralism is not committed to saying that Sally has this belief. An egoist has a shorter list of ultimate desires than a pluralist, but the egoist has a longer list of causal beliefs. For this reason, it is unclear why psychological egoism should be regarded as the more parsimonious theory (Sober and Wilson 1998).

An Evolutionary Approach

Psychological motives are proximate mechanisms in the sense of that term used in evolutionary biology. When a sunflower turns towards the sun, there must be some mechanism inside the sunflower that causes it to do so. Hence, if phototropism is an adaptation that evolved because it provided organisms with certain benefits, then a proximate mechanism that causes that behavior also must have evolved. Similarly, if certain forms of helping behavior in human beings are evolutionary adaptations, then the motives that cause those behaviors in individual human beings must also have evolved. Perhaps a general perspective on the evolution of proximate mechanisms can throw light on the specific problem of whether egoism or motivational pluralism was more likely to have evolved.

Pursuing this evolutionary approach does not presuppose that every detail of human behavior, or every act of helping, can be completely explained by the hypothesis of evolution by natural selection. Doubtless there are many facts about behavior and many instances of helping for which natural selection is not a relevant explanation. However, I want to consider a single fact about human behavior, and my claim is that selection is relevant to explaining it. The phenomenon of interest is that human parents take care of their children; the average amount of parental care provided by human beings is strikingly greater than that provided by parents in many other species. I will assume that natural selection is at least part of the explanation of why parental care evolved in our lineage. This is not to deny that human parents vary; some parents take better care of their children than others, and some even abuse and kill their offspring.

It also may be true that mothers, on average, do more to take care of their children than fathers do, though here it is important to remember that

provisioning is a form of parental care just as nursing is. In any event, the question of how and why the sexes differ in their contributions to parental care is not my subject.

To tease out some general principles that govern how one might predict the proximate mechanism that will evolve to cause a particular behavior, I will switch examples to a hypothetical mindless organism whose problem is to select items from its environment to eat. Some particles that float by in the liquid medium in which the organism lives contain protein; others contain poison. The organism has evolved a particular behavior – it tends to eat protein and avoid poison. What proximate mechanism might have evolved that allows it to do so?

First let us survey the range of possible design solutions that we need to consider. The most obvious design solution to this problem is for the organism to have a detector that distinguishes protein from poison. It captures a morsel that floats by, puts the particle in its detector, and then has the output of this detector wired to a behavior; the organism either eats the morsel or spits it out. I will call this the *direct* solution to the design problem; the organism needs to discriminate between protein and poison, and this solution accomplishes that end by using a detector that detects that very contrast in properties.

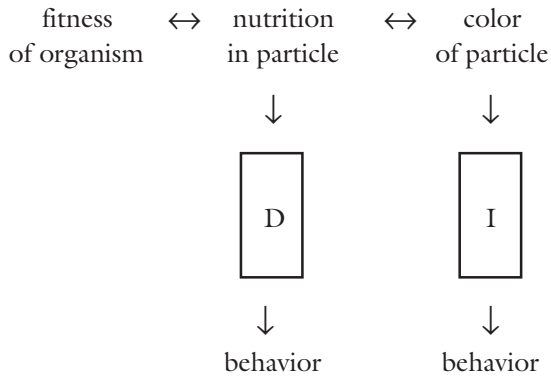
It is not hard to imagine other solutions to the design problem that are less direct. Suppose that protein tends to be red and that poison tends to be green in the organism's environment. If so, the organism could use a color detector to make the requisite discrimination. This design solution is *indirect*; the organism needs to distinguish protein from poison and accomplishes this by discriminating between two other properties that happen to be correlated with the target contrast. In general, there may be many indirect design solutions that the organism might exploit; there are as many indirect solutions as there are correlations between the protein/poison distinction and other properties found in the environment. Finally, we may add to our list the idea that there can be pluralistic solutions to a design problem. In addition to the monistic solution of having a protein detector and the monistic solution of having a color detector, an organism might deploy both a protein detector *and* a color detector.

Given this multitude of possibilities, how might one predict which of them will evolve? Three principles are relevant here – *availability*, *reliability*, and *efficiency* (Sober 1994; Sober and Wilson 1998).

Natural selection acts only on the range of variation that exists ancestrally. A protein detector might be a good thing for the organism to have, but if that device was never present as an ancestral variant, then natural selection cannot cause that trait to evolve. So the first sort of information we would like to have concerns which proximate mechanisms were *available* ancestrally.

Let us suppose for the sake of argument that both a protein detector and a color detector are available ancestrally. Which of them is more likely to evolve? Here we need to address the issue of *reliability*. Which device does the more reliable job of indicating which particles in the environment are good to eat? Without further information, not much can be said. A color detector may have any degree

Figure 7.2



of reliability, and the same is true of a protein detector. There is no a priori reason why the direct strategy should be more or less reliable than the indirect strategy. However, there is a special circumstance in which they will differ. It is illustrated by Figure 7.2. The double arrows in Figure 7.2 indicate correlation; gaining nutrition is correlated with an organism's fitness, and a particle's being red rather than green is correlated with its nutritional content. There is no sequence of arrows from fitness to color except the one that passes through nutrition. This means that an organism's fitness is correlated with the color of the particles that it eats. There is no a priori reason that color should be relevant to fitness only by virtue of indicating nutritional content. For example, if eating red particles attracted predators more than eating green ones does, then color would have two sorts of relevance for fitness. However, if nutrition "screens off" fitness from color in the way indicated in the figure, we can state the following principle about the reliability of the direct device D and the indirect device I:

(D/I) If nutrition and color are less than perfectly correlated, and if D detects nutrition at least as well as I detects color, then D will be more reliable than I.

This is the Direct/Indirect Asymmetry Principle. Direct solutions to a design problem are not always more reliable, but they are more reliable in the circumstance described.

A second principle about reliability also can be extracted from Figure 7.2. Just as scientists do a better job discriminating between hypotheses if they have more evidence rather than less, so it will be true that organisms make more reliable discriminations if they have two sources of information about what to eat rather than just one:

(TBO) If nutrition and color are less than perfectly correlated, and if D and I are each reliable, though fallible, detectors of nutrition, then D

and I working together will be more reliable than either of them working alone.

This is the Two-is-Better-than-One Principle. It requires an assumption that the two devices do not interfere with each other when they are both present in an organism; they function fairly independently.

The D/I Asymmetry and the TBO Principle pertain to the issue of reliability. Let us now turn to the third consideration that is relevant to predicting which proximate mechanism will evolve, namely *efficiency*. Even if a nutrition detector and a color detector are both available, and even if the nutrition detector is more reliable, it does not follow that natural selection will favor the nutrition detector. It may be that a nutrition detector requires more energy to build and maintain than a color detector. Organisms run on energy no less than automobiles do. Efficiency is relevant to a trait's overall fitness just as much as its reliability is.

With these three considerations in hand, let us return to the problem of predicting which motivational mechanism for providing parental care is likely to have evolved in the lineage leading to present-day human beings. The three motivational mechanisms we need to consider correspond to three different rules for selecting a behavior in the light of what one believes:

(HED) Provide parental care if and only if doing so will maximize pleasure and minimize pain.

(ALT) Provide parental care if and only if doing so will advance the welfare of one's children.

(PLUR) Provide parental care if and only if doing so will either maximize pleasure and minimize pain, or will advance the welfare of one's children.

(ALT) is a relatively direct, and (HED) a relatively indirect, solution to the design problem of getting an organism to take care of its offspring. Just as an organism can find nutrition by detecting color, so it is possible in principle for a hedonistic organism to be built in such a way that it will provide parental care; what is required is that the organism be so constituted that providing parental care is the thing that maximizes its pleasure and minimizes its pain (or that the organism at least believes that this is so).

Let us consider how reliable these three mechanisms will be in a certain situation. Suppose that a parent learns that its child is in danger. Imagine that your neighbor tells you that your child has just fallen through the ice on a frozen lake. Figure 7.3 shows how (HED) and (ALT) will do their work. The altruistic parent will be moved to action just by virtue of believing that its child needs help. The hedonistic parent will not; rather, what moves the hedonistic parent to action is the feelings of anxiety and fear that are caused by the news, or the parent's belief that such negative feelings will continue unless the child's situation is improved. It should be clear from Figure 7.3 that the (D/I) Asymmetry Principle applies. In the circumstance specified, (ALT) will be more reliable than (HED). And

Figure 7.3

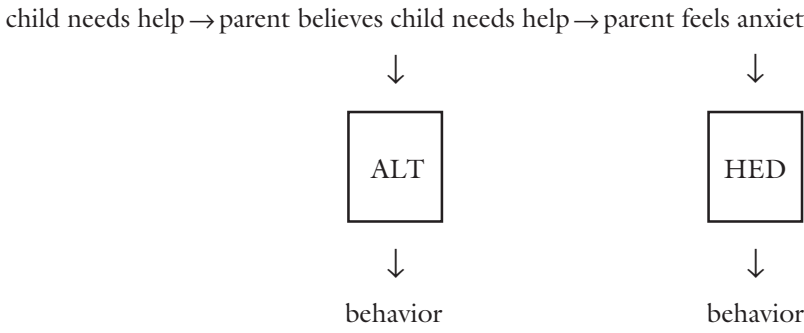
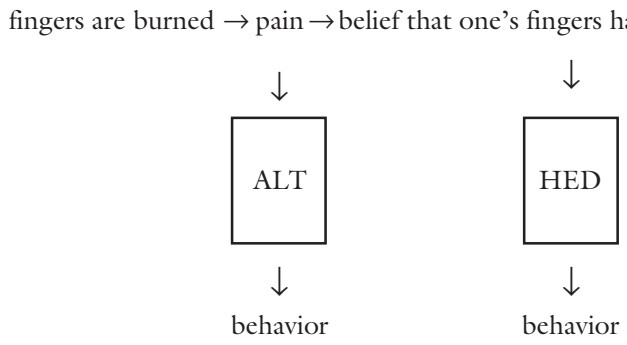


Figure 7.4



by the (TBO) Principle, (PLUR) will do better than both. In this example, hedonism comes in last in the three-way competition, at least as far as reliability is concerned.

The important thing about this example is that the feelings that the parent has are *belief mediated*. The only reason the parent feels anxiety and fear is that the parent *believes* that its child is in trouble. This is true of many of the situations that egoism and hedonism are called upon to explain, but it is not true of all of them. For example, consider the situation in Figure 7.4 in which pain is a direct effect and belief a relatively indirect effect, of bodily injury. Now hedonism is a direct solution to the design problem; it would be silly to build the organism so that it is unresponsive to pain and withdraws its fingers from the flame only after it forms a belief about bodily injury. In this situation, *belief is pain-mediated* and the (D/I) Asymmetry Principle explains why a hedonistic focus on pain makes sense. However, the same principle indicates what is misguided about hedonism

as a design solution when pain is belief-mediated, which is what occurs so often in the context of parental care.

If hedonism is less reliable than either pure altruism or motivational pluralism, how do these three mechanisms compare when we consider the issues of evolutionary availability and efficiency? With respect to availability, I want to make this claim: *if hedonism was available ancestrally as a design solution, so was altruism*. The reason is that the two motivational mechanisms differ in only a very minor way. Both require a belief/desire psychology. And both the hedonistic and the altruistic parent want their children to do well; the only difference is that the hedonist has this propositional content as an instrumental desire while the altruist has it as an ultimate desire. If altruism and pluralism did not evolve, this was not because they were unavailable as variants for selection to act upon.

What about the question of efficiency? Does it cost more calories to build and maintain an altruistic or a pluralistic organism than it does to build and maintain a hedonist? I do not see why. What requires energy is building the hardware that implements a belief/desire psychology. However, it is not easy to see why having one ultimate desire rather than two should make an energetic difference; nor is it easy to see why having the ultimate desire that your children do well should burn more calories than having the ultimate desire to avoid pain and attain pleasure. People with more beliefs apparently do not need to eat more than people with fewer. The same point seems to apply to the issue of how many, or which, ultimate desires one has.

In summary, hedonism is a less reliable mechanism than pure altruism or pluralism as a device for delivering parental care. And, with respect to the issues of availability and efficiency, we found no difference among these three motivational mechanisms. This suggests that natural selection is more likely to have made us motivational pluralists than to have made us hedonists.

From an evolutionary point of view, hedonism is a bizarre motivational mechanism. What matters in the process of natural selection is an organism's ability to survive and be reproductively successful. Reproductive success involves not just the production of offspring, but the survival of those offspring to reproductive age. What matters in the process of natural selection is the survival of one's own body and the bodies of one's children. Hedonism, in contrast, says that organisms care ultimately about the states of their own consciousness, and about that alone. Why would natural selection have led organisms to care about something that is peripheral to fitness, rather than have them set their eyes on the prize? If organisms were unable to conceptualize propositions about their own survival and the production and survival of their offspring, that might be a reason. After all, it can make sense for an organism to exploit the indirect strategy of deciding what to eat on the basis of color rather than on the basis of nutritional value, if the organism has no direct access to nutritional content. But if an organism is smart enough to form representations about itself and its offspring, this justification of the indirect strategy will not be plausible. The fact that we evolved from ancestors who were cognitively less sophisticated makes it unsurprising that avoiding pain and

attaining pleasure are two of our ultimate goals. But the fact that human beings are able to form representations with so many different propositional contents suggests that evolution supplemented this list of what our ultimate ends are (see Stich 2007 and Schulz 2011 for further discussion).

Concluding Comments

I have argued that past philosophical and psychological attempts to resolve the debate between egoism and motivational pluralism have not succeeded. It would be astonishing if this dispute about an apparently empirical matter could be resolved by arguments a priori. Unfortunately, the observations that people casually make in ordinary life and that scientists make in the laboratory have not been decisive either; although some simple versions of egoism are refuted by what we observe, other versions of egoism can be constructed that seem to fit the available observations. Perhaps more sophisticated experiments and observations of behavior will answer the question. But for now, the situation in philosophy and psychology is one of stalemate.

Can evolutionary considerations break through this impasse? The argument of the previous section aims to establish that a purely egoistic set of motives is less likely to have evolved than a set of motives that includes both egoistic and altruistic ultimate desires. I do not suggest that this argument *proves* that people are motivational pluralists; there is much that remains unknown about the mind and how it evolved, and there is no guarantee that further details will not substantially alter the picture I have tried to develop. However, I do think that the argument suffices to show that egoism does not deserve to be regarded as the default hypothesis that we should accept as long as it is consistent with what we observe. In my opinion, the weight of evidence favors pluralism, if only to a small degree.

References

Batson, C.D. (1991) *The Altruism Question: Toward A Social-Psychological Answer*, Hillsdale, NJ: Lawrence Erlbaum Associates.

Broad, C.D. (1965) *Five Types of Ethical Theory*, Totowa, NJ: Littlefield, Adams.

Butler, J. (1726/1965) *Fifteen Sermons upon Human Nature*, reprinted in *British Moralists: Being Selections From Writers Principally of the Eighteenth Century*, vol. 1, ed. L.A. Selby-Bigge, New York: Dover Books, pp. 180–241.

Clark, R.D. and Word, L.E. (1974) "Where is the Apathetic Bystander? Situational Characteristics of the Emergency," *Journal of Personality and Social Psychology* 29: 279–87.

Feinberg, J. (1984) "Psychological Egoism," in *Reason at Work*, eds. S. Cahn, P. Kitcher, and G. Sher, San Diego, CA: Harcourt Brace and Jovanovich, pp. 25–35.

Gibbard, A. (1990) *Wise Choices, Apt Feelings*, Cambridge, MA: Harvard University Press.
Hume, D. (1751/1970) "On Self Love," in *An Enquiry Concerning the Principles of Morals*, Indianapolis: Hackett.
Kavka, G. (1986) *Hobbesian Moral and Political Theory*, Princeton, NJ: Princeton University Press.
LaFollette, H. (1988) "The Truth in Psychological Egoism," in *Reason and Responsibility*, 7th edn, ed. J. Feinberg, Belmont, CA: Wadsworth, pp. 500–7.
Nagel, T. (1970) *The Possibility of Altruism*, Oxford: Oxford University Press.
Nozick, R. (1974) *Anarchy, State, and Utopia*, New York: Basic Books.
Schlick, M. (1939) *Problems of Ethics*, New York: Prentice Hall.
Schroeder, D., Penner, L., Dovidio, J., and Piliavin, J. (1995) *The Psychology of Helping and Altruism*, New York: McGraw-Hill.
Schulz, A. (2011) "Sober & Wilson's Evolutionary Arguments for Psychological Altruism – a Reassessment," *Biology and Philosophy* 26: 251–60.
Sidgwick, H. (1907/1922) *The Methods of Ethics*, 7th edn, London: Macmillan.
Sober, E. (1992) "Hedonism and Butler's Stone," *Ethics* 103: 97–103.
Sober, E. (1994) "Did Evolution Make Us Psychological Egoists?" in *From a Biological Point of View: Essays in Evolutionary Philosophy*, New York: Cambridge University Press, pp. 8–27.
Sober, E. and Wilson, D.S. (1998) *Unto Others – The Evolution and Psychology of Unselfish Behavior*, Cambridge, MA: Harvard University Press.
Stampe, D. (1994) "Desire," in *A Companion to the Philosophy of Mind*, ed. S. Guttenplan, Oxford: Blackwell, pp. 244–50.
Stewart, R.M. (1992) "Butler's Argument Against Psychological Hedonism," *Canadian Journal of Philosophy* 22: 211–21.
Stich, S. (2007) "Evolution, Altruism and Cognitive Architecture: A Critique of Sober and Wilson's Argument for Psychological Altruism," *Biology and Philosophy* 22: 267–81.

Chapter 8

The Science of Ethics

Ron Mallon and John M. Doris

> Perhaps the light will prove another tyranny. Who knows what new things it will expose?
>
> C.P. Cavafy, The Windows

The Science Wars

Perhaps the most visible trend in philosophical ethics over the first years of the twenty-first century has been the remarkable number of moral philosophers referencing, and producing, empirical work. In moral psychology and experimental philosophy, the fields where this "empirical turn" is most evident, papers, anthologies, and monographs are appearing at a dizzying clip.[1] Among the philosophers and scientists involved, the tone is often exuberant, with partisans claiming progress in debates that have been ossified for many a year (e.g., Stich, Doris, and Roedder 2010: 202).

For those who are not so strongly identified with the empirical turn, the mood is less celebratory. Daniel Jacobson, principal investigator on a $1.2 million Templeton Foundation grant to study "The Science of Ethics," contends that "too often advocates of the empirical ethics movement overreach in their conclusions, in ways that beg the most important philosophical questions and threaten the very possibility of moral reasoning."[2] One shudders to think what people who are not getting $1.2 million to study "the science of ethics" are saying about it. As well one might: in the popular media, experimental philosophy has been called an embarrassment, and its proponents have been said to hate philosophy.[3] We will

The Blackwell Guide to Ethical Theory, Second Edition. Edited by Hugh LaFollette and Ingmar Persson.

not attempt a global diagnosis of the vitriol directed at empirically informed philosophical work, but we do think it is worth considering the anxiety it inspires within philosophical ethics, even among those who aim to produce it.

There is more than one anxiety in the air. First, there are charges of "overreaching." The tendency to draw grand conclusions from the one or two studies that best suits one's purposes – "cherry-picking," as some psychologists say – is, most assuredly, to be avoided: if there are philosophical lessons in science, they emerge through consideration of robust research traditions, not isolated experiments. But we doubt that empirically minded philosophers are especially prone to cherry-picking: they love to cite chapter and verse of empirical studies, and often do so till the cows come home – sometimes even longer. In our experience, it is just as likely to be their opponents who overreach, alleging that difficulty with an experiment or two is sufficient to scuttle an empirically based approach (Doris 2005: 660, fn. 10).

More generally, we find worries about empirically inspired overreaching a little puzzling. Are the theoretical creations of empirically oriented philosophers more vaultingly ambitious than idealism, monism, or possible worlds realism? If empirical philosophers sometimes overreach the data (as do scientists themselves), many philosophers do their reaching *without any data*. In our estimation, a distinctive feature of philosophical inquiry *just is* overreaching, and that is a *good* thing; if there is a complaint about reaching empirical methods, it is more likely that the empirically oriented do not reach *far enough*, and let the dense weight of fact crush the air from philosophical imagination. Philosophical method has always been the method of agonistic dialectic: reach till your knuckles are rapped, retrench, and reach again. Philosophy is better – and more fun – for it.

A second anxiety concerns "question-begging." This might be – there is lots of that around. If the old saw is to be believed, all philosophical arguments are either ad hominem or question-begging; assuming both forms are equally represented, that is a multitude of begged questions. But it is not easy to determine when a question has been begged, in part because it is not easy to precisely explicate what the fallacy is supposed to be (for able discussion, see Sorenson 1991, 1996, 1999). Indeed, it is not entirely clear that begging the question *is* a fallacy, since what looks to be a paradigmatic form of begging the question, "*P*, therefore *P*," is a valid argument. Often enough, when question-begging is charged, the complaint is that the defendant is assuming what is at issue. But trying this charge is unlikely to be a matter of identifying a straightforward fallacy. When it comes to that, we suspect that the Boulevard of Begged Questions is a two-way street, where each contesting party can lodge such a complaint against the other with equal merit.

This circumstance becomes evident with a (somewhat) concrete example. While it is not entirely clear what "philosophical naturalism" is supposed to be, it is pretty clear that many empirical types, with their affection for science, subscribe to this doctrine – as we ourselves do. A standard complaint about naturalists, at least in

ethics, is that they cannot make appropriate sense of normativity (more quaintly: they disrespect the is/ought or fact/value distinction). For this charge to be intelligible there has to be an operative understanding of normativity: before you say *we* cannot make sense of it, *you* had better say what it is. But the appropriate understanding of normativity is precisely one of the things at issue, and the naturalist will likely want to contest the critic's understanding. That is, for the nonnaturalist to file her charge, she has to begin by asserting something the naturalist begins by rejecting: she asserts what is at issue. To be sure, the naturalist is no less guilty of this than the nonnaturalist: each side proceeds with assumptions that the other rejects. As we said, we are all driving a two-way street.

A third anxiety, and the anxiety that this essay is meant to assuage, concerns a sort of moral nihilism: the suspicion that empirical approaches to ethics "threaten the very possibility of moral reasoning." Of course, the empirical work does not show that nobody reasons about moral matters, but it might be taken to show that nobody reasons *well* about moral matters. For example, it might be taken to show that the taint of bias is impossible to escape. This would be an unhappy state of affairs. But that does not mean it is a state of affairs the existence of which can be ruled out prior to extensive empirical investigation. If anthropogenic climate change is real, and has catastrophic consequences, it makes depressing news, but the fact that this news is depressing makes a lousy justification for climate change denialism. Hard truths are, after all, truths, and the truth about human moral capacities, like the truth about environmental degradation, might be hard.

If science showed that moral reasoning – or rather good moral reasoning – is psychologically impossible, that would be a hard truth. However, we doubt it is a truth. To be sure, much empirical work on morality has taken a skeptical tone, casting doubt on familiar materials of ethical thought, such as character (Doris 2002), intuition (Sinnott-Armstrong 2007d), and rationality (Stich 1990). And numerous prominent proponents of empirical approaches are drawn to moral irrealism (e.g., Greene 2007; Nichols 2004; Prinz 2007). Perhaps much of morality, which dates to a prescientific worldview, will wither in the continued light of science. Or perhaps not. A defensible verdict – just as verdicts on charges of overreaching or question-begging – can only be reached on a case-by-case basis, via detailed consideration of the relevant empirical evidence, philosophical positions, and their points of interaction. It is this sort of consideration we begin to undertake here.

Why might the empirical study of morality lead to moral nihilism? As with any complex social phenomenon there are, no doubt, many causes, but our diagnosis will focus on one. A central trend in the cognitive and social psychology of the past thirty years has been an increasing appreciation of the role that unconscious and automatic processes play in the production of cognitive states and behaviors. In the study of morality, this has meant a shift away from attempts to characterize moral judgment as a product of reasoning (e.g., Kohlberg 1969), or as a sort of

information processing (e.g., Darley and Schultz 1990), in favor of *psychological intuitionist* accounts that emphasize a dominant role for unconscious, intuitive processes, especially emotions, in moral judgment (e.g., Blair 1995; Haidt 2001; Greene *et al.* 2001; Greene 2007 cf. Mallon and Nichols 2010; Maibom 2010). This trend towards accounts that insist on a limited role for reflection, deliberation, and reasoning in moral functioning is, we submit, the feature of recent empirical work that seems so threatening to traditional conceptions of our moral capacities: the moral animal, it appears, is an irrational animal.

We will take this unsettling body of research as our starting place, and consider some of the skeptical implications that have been claimed for it. While we believe these findings trouble important aspects of commonsense (and common philosophical) moral thinking (Doris 2009, forthcoming), we doubt there is an easy path to skeptical conclusions about moral reasoning or judgment. Ethical theorists need to take the research seriously, but at this point, there is little reason to fear doing so will result in self-immolation.

The therapy we propose for easing this nihilistic anxiety takes the following course: In the section "From Psychological Intuitionism to Skepticism," we outline psychological intuitionist accounts of moral judgment, and note some ways in which they seem to pose a threat to successful moral reasoning.

In the section "Debunking Arguments," we will consider some implications of psychological intuitionism. Here, we will provisionally concede the dominance of automatic processes and consider *debunking arguments* suggesting that moral judgments are sourced in automatic processes that are not responsive to relevant moral considerations (Greene 2007, forthcoming; Joyce 2006; Singer 2005; cf. Mason 2011). We will argue that these arguments depend on an excessively narrow conception of intuitive processes; with a more liberal – and plausible – understanding, the force of these arguments is blunted.

In the section "Are Intuitions Dominant?" we will rescind our provisional concession of the dominance of automatic processes, and challenge psychological intuitionism on the grounds that the experimental evidence adduced for it does not establish the dominance of intuition. We consider one argument for such psychological intuitionism, the *argument from finite resources* (Mallon and Nichols 2011), and suggest that it is not decisive, since there are alternative models of moral judgment that are untroubled by it.

In the section "Epiphenomenal Moral Reasoning," we will consider another argument that seems to threaten ordinary moral reasoning: the *argument from the epiphenomenality*. This argument holds that moral reasoning is usually, or always, causally inert, meaning it cannot figure in explanation of moral judgments. We will suggest that this argument trades on a conflation of explanation and justification.

We will close, in the section "Culture and Moral Life," with a somewhat speculative diagnosis of where psychological intuitionism goes wrong: its failure to adequately account for the role of individual experience and transmitted culture in fixing the content of moral judgment.

From Psychological Intuitionism to Skepticism

The last fifteen years of work in empirically informed moral psychology has effected a pronounced shift towards characterizing the mechanisms of moral judgment and other moral behaviors in "dual-process" terms. While different accounts construe these features somewhat differently, the central idea is fairly clear. On the one hand are "system 1" or "intuitive" processes that are phylogenetically ancient, fast, unconscious, and effortless. On the other hand are "system 2" or "reasoning" processes that are phylogenetically recent, slow, conscious, and effortful (Haidt 2001; Greene *et al.* 2001; cf. Stanovich 2004; Bargh and Chartrand 1999; Chaiken and Trope 1999; Evans and Frankish 2009). For us, nothing much turns on precise characterization of the dichotomy, or whether or not it is exhaustive (a dualism we would find surprising). Here the crucial issue concerns a hypothesis often associated with (though not necessarily entailed by) dual-process accounts: system 1 processes play a dominant role in the production of moral judgment. Call this the *dominance hypothesis*:

> *The dominance hypothesis*: system 1 processes typically determine moral judgment.

According to such accounts, it is typically or normally the system 1 intuitive processes, rather than the system 2 reasoning processes, that explain why people make the moral judgments they do.[4]

We can see an emerging recognition of the role of emotion in human judgment in the work of R. James Blair. Inspired by Konrad Lorenz's suggestion that social animals possess a mechanism to modulate intraspecies or intragroup aggression, Blair suggested that humans typically possess an automatic, emotional response to moral violations underwritten by a "violence inhibition mechanism" or "VIM" that functions to cause us to withdraw from aggression when we witness low-level sensory cues indicating the distress of a conspecific. Blair goes on to argue that VIM is intact in autistic people (Blair 1996; cf. Leslie, Mallon, and DiCorcia 2006) and impaired in psychopaths (Blair 1995; Blair, Mitchell, and Blair 2005), giving rise to the inability of the latter, but not the former, to distinguish moral from conventional violations. In other words, a normally functioning VIM is necessary for facility with standard moral judgments. Since it is spared in autistic people, they perform pretty normally on this score (despite other impairments in social reasoning), while the compromise of the VIM in psychopaths impairs moral reasoning (despite their oft-observed facility in social interaction). For Blair, then, a central sort of moral judgment is subserved by an emotional response that operates automatically when triggered by cues to an evolutionarily important situation (though it later becomes paired with more sophisticated moral responses).

Emotion becomes more fully and explicitly married to dual-process accounts of cognition in Jonathan Haidt's work on moral judgment (2001). For Haidt,

system 1 processes are "intuitions," of which emotions are an important subset, and system 2 processes are "reasoning." He writes:

> The words *intuition* and *reasoning* are intended to capture the contrast made by dozens of philosophers and psychologists between two kinds of cognition. . . . intuition occurs quickly, effortlessly, and automatically, such that the outcome but not the process is accessible to consciousness, whereas reasoning occurs more slowly, requires some effort, and involves at least some steps that are accessible to consciousness.
>
> (2001: 818)

Haidt's influential application of dual-process thinking to moral judgment provides the background against which assertions of the dominance of intuition by Haidt and others make sense.

Joshua Greene also uses a dual-process account of cognition as a theoretical guide for his extraordinary neuroscientific investigations of morality (e.g., Greene *et al.* 2001; Greene *et al.* 2008; Greene forthcoming). Greene and his colleagues (2001) follow others (Petrinovich, O'Neill, and Jorgensen 1993; O'Neill and Petrinovich 1998; Mikhail 2000) in using classic philosophical moral dilemmas developed by Philippa Foot and Judith Jarvis Thompson as the basis for empirical investigations that illuminate the properties relevant to our moral responses. Perhaps the most famous of these moral dilemmas concerns a runaway trolley. In the *Switch Case*, subjects imagine that a person sees a trolley approaching that will kill five innocents on the track, and the only way to prevent the deaths of these five is to flip a switch that will divert the train to a side track, resulting in the death of one person there. Philosophers have maintained that common sense endorses flipping the switch to divert the train, leading to the death of one instead of five (e.g., Thomson 1976). In the *Footbridge Case*, the situation is similar except that the only way to save the five is to push a (very) large stranger off of a footbridge in front of the oncoming train, which will kill the stranger but, by virtue of his or her impressive mass, stop the train and save the five. In this case, philosophers have suggested that the commonsense position is that it is wrong to push the stranger (e.g., Foot 1978; Quinn 1989; Thomson 1976). Yet on a simple utilitarian calculus, the cases are parallel: five lives can be saved at a cost of one.

Greene proposed to explain this tension with the suggestion that the response in footbridge-style cases is generated by the fact that pushing a person is "personal," resulting in greater emotional engagement than switch-style "impersonal" actions.[5] He characterizes the distinction as follows:

> A moral violation is personal if it is: (i) likely to cause serious bodily harm, (ii) to a particular person, (iii) in such a way that the harm does not result from the deflection of an existing threat onto a different party . . . A moral violation is impersonal if it fails to meet these criteria. . . . Pushing someone in front of a trolley meets all three criteria and is therefore "personal," while diverting a trolley involves merely deflecting

> an existing threat, removing a crucial sense of 'agency' and therefore making this violation "impersonal."
>
> (Greene and Haidt 2002: 519)

According to Greene, impersonal dilemmas tend to activate a system 2 "reason" path, whereas personal dilemmas run through the system 1 "emotion" path. So, like Haidt, Greene's model suggests that we can arrive at moral judgments either through system 2, reasoning processes, or through system 1, intuitive processes. However, both Greene and Haidt go beyond merely asserting an important role for intuitive processes to suggesting that system 1 processes are dominant. In a coauthored article, they write: "Moral judgment is more a matter of emotion and affective intuition than deliberate reasoning," suggesting that "we see an action or hear a story and we have an instant feeling of approval or disapproval" that appears "suddenly and effortlessly in consciousness" (Greene and Haidt 2002: 517). It is this view, pairing a dual-process account of cognitive processes with the dominance hypothesis, that we characterize as "psychological intuitionism."

Of course, it seems that people very often (in both ordinary social life and academic philosophy) reason about what is morally appropriate. Greene and Haidt suggest that this appearance is misleading: both maintain that reasoning about the moral domain is typically only a means for *rationalizing* our antecedently produced moral judgments or other behaviors. Perhaps the best-known illustration of this model is Haidt's notorious vignette:

> Julie and Mark are brother and sister. They are traveling together in France on summer vacation from college. One night they are staying alone in a cabin near the beach. They decide that it would be interesting and fun if they tried making love. At the very least, it would be a new experience for each of them. Julie was already taking birth control pills, but Mark uses a condom too, just to be safe. They both enjoy making love, but they decide not to do it again. They keep that night as a special secret, which makes them feel even closer to each other. What do you think about that? Was it okay for them to make love?
>
> (2001: 814)

Haidt reports that subjects presented with such cases and asked for a moral judgment typically answer "immediately" that the action is wrong, and *then* begin "searching for reasons" to support their judgment. This seems to illustrate Haidt's idea that system 1 intuitions generate the initial response, only to be followed by system 2 reasoning processes. Moreover, Haidt and colleagues designed the case to deprive subjects of the most obvious justifications for their judgment. For example, a subject might try to explain the judgment that the act was wrong by appeal to the emotional harm incest would cause, only to be answered by the researcher noting to them that (according to the vignette) Mark and Julie suffered no harm at all (Haidt, Bjorklund, and Murphy n.d.). The result of this process is that many subjects find themselves unable to provide reasons that justify their

judgment. Nevertheless, they continue to insist that the judgment is correct. For Haidt, the failure of these "dumbfounded" subjects to produce successful justifications for their moral judgments is further evidence that reasoning plays little role in generating the judgments.

Haidt's suggestion conforms with a host of experimental evidence (e.g., Nisbett and Wilson 1977; Wilson 2002; Stanovich 2004) to the effect that people (at least sometimes) are not consciously aware of the processes that connect their mental states to the causal effects of those states, and that people routinely confabulate explanations for their behavior. In a diabolic stroke, Greene (2007) goes even further, suggesting that Kantian moral theories are themselves elaborate rationalizations of system 1 emotional responses. For instance, Kantian objections to pushing the large man in front of the trolley in a footbridge case on the grounds that to do so is to use the man as a mere means are, on Greene's view, elaborate rationalizations of a simple emotional aversion to inflicting (something like) "up close and personal" harms.

In sum, the recent turn to psychological intuitionism has resulted in a model with three features that seem to invite skepticism about moral reasoning.

First, the model holds that system 1 automatic processes, rather than system 2 reasoning processes, are the primary psychological determinant of moral judgment and behavior. For those who hold reason itself to be essential to moral judgment or behavior, this dominance hypothesis entails that we rarely engage in genuine moral judgment or behavior (Kennett and Fine 2009).

Second, system 1 processes are conceived so as to be not conducive to apprehending morally relevant considerations. This could be because automatic processes are (1) sensitive to factors that have nothing to do with morality, or (2) insensitive to factors relevant to moral judgment, or (3) fail to meet some more general requirements of good reasoning – for example, fail to result in coherent judgments. While there are many specific studies intimating that one or more of these conditions sometimes obtains, there is a compelling reason for believing they might obtain generally; namely, system 1 processes are phylogenetically ancient, automatic processes that are designed to respond to properties (of situations, actions, events, or people) that were important to our ancestors' survival and reproduction – properties of evolutionary import. But there is no obvious reason to think that these properties of evolutionary import neatly overlap (or even overlap significantly) with morally relevant properties.[6]

The third salient feature of psychological intuitionism is the suggestion that *moral reasoning* is typically epiphenomenal with regard to the production of moral judgment and behavior. "Moral reasoning" here can be understood either as a type of cognition (underwritten largely by system 2 mechanisms) or as a type of behavior manifested either individually or in concert with others. Either way, psychological intuitionists hold that moral reasoning is not typically the cause of moral thought and behavior.

Together these three features of the model might be taken to suggest that humans are typically morally incompetent. People's moral responses are generated

by mechanisms that are insensitive to morality and highly sensitive to morally irrelevant considerations; after these suspect judgments are reached, people compound their errors by rationalizing their responses.

Debunking Arguments

Let us provisionally concede that the dominance hypothesis is correct. Should this lead us to skepticism about moral judgments? One influential *debunking* argument says "yes," insisting that the (automatic) mechanisms that give rise to our moral responses fail to be reliable guides to morality.

Why think this? While a number of arguments have been proposed, we will focus on one closely connected to an emphasis on system 1 processing.[7] The argument grows out of evolutionary psychological theorizing about the emotions (e.g., Lazarus 1994) and out of other parts of evolutionary psychology that emphasize the adaptedness of our psychological processes solving problems in our ancestral environment to our ancestors' successful survival and reproduction. For Blair, this could be the distress of a conspecific, while for Greene, this could be some sort of "up close and personal" contact.[8] The argument proceeds by suggesting that once we understand what the mechanism is really tracking, it undermines our faith that it is responding to morally relevant features of the world (Greene 2007, forthcoming; Singer 2005; Joyce 2006). As Joyce colorfully puts it: "We should reject or modify any theory that would render us epistemic slaves to the baby-bearing capacity of our ancestors" (2006: 219). The result is a debunking of some or all of our moral responses. From here, some theorists are tempted to reconstruct a moral view on other grounds (Singer 2005), others seem tempted by moral skepticism (Joyce 2006), and still others opt for a bit of both (Greene forthcoming).

Thus conceived, moral debunking arguments have a dual aspect. One part is a *causal explanation* of our moral responses while the other part is a *moral gap* argument to the effect that the cause thus understood gives us no reason to endorse the moral content of these responses.

Greene (2007) provides a nice elucidation of results that make a moral gap argument look promising, identifying a number of properties that the empirical evidence suggests are *psychologically* relevant in moral judgment and behavior:

- People view harm caused by "up close and personal" bodily contact to be worse than harm caused at a distance.
- People view moral obligation to suffering persons who are near to be greater than obligation to those far away.
- People view a harm as greater if perpetrated against an identifiable victim as opposed to a generic victim.
- People elect to punish concrete individuals more than abstract individuals.

For Greene these are products of unconscious mechanisms that are themselves adaptations for achieving evolutionary advantage. But crucially, these examples suggest a moral gap: our judgments are responsive to many properties that are (at least prima facie) not *morally* relevant (Greene 2003). The system 1 explanation paired with evidence for a moral gap gives rise to the sense that our moral responses (or at least some of them) are being debunked (cf. Singer 2005; Mason 2011).

However, debunking moral responses requires more than showing that some moral responses are sometimes sensitive to morally irrelevant properties. It requires showing that our moral psychology, or parts of it, does not in general reliably track morally relevant properties. And this, as a number of commentators have noted, is not easily shown.

Automatic and emotional processes look to be implicated in normal, successful moral cognition (Blair 1995; Damasio 1994; Mason 2011; Narvaez 2010). Emotional deficits in psychopaths (Blair 1995; Blair, Mitchell, and Blair 2005) also suggest that emotional impairment leads to moral impairment (but see Maibom 2005).[9] Now, perhaps these arguments are mistaken and can be answered. The present point is that even if we assume intuitive processes are dominant, we still lack evidence that they result in moral error often enough to be unreliable.

What could fill in this argument? Specific studies indicating that moral judgment is sensitive to one or another irrelevant properties run afoul of something analogous to what Mallon and Nichols (2011) have called *the counting problem*; even if we allow that the study identifies a process plausibly thought to be found in natural contexts, we still have to wonder *how frequent* the failing is in everyday life.[10] What is needed is some general reason to think that everyday moral reasoning is typically like the cases in the lab (at least, if we are willing to concede the dominance of system 1 processes). A plausible way of filling in the argument is with the insistence that the intuitive processes underlying our moral judgments are both inflexible and adapted to our evolutionary circumstances rather than our present ones (Greene 2003, 2007, forthcoming; Joyce 2006; Singer 2005). If this were true, then it would give us a general argument for believing there is a moral gap.

In fact, the inflexibility of system 1 processes is, on the standard story, part of their functional adaptiveness. As Greene puts it, "When Nature needs to get a behavioral job done, it does it with intuition and emotion wherever it can" (Greene 2007: 60). Apparently, the tendency is to automate solutions to any problems that are stable enough to be solved by automatic processing, leaving the precious resources of conscious reasoning to address those problems that cannot be automated. And there is considerable evidence for a certain kind of inflexibility: intuitive, automatic processes, including emotions, often seem to have activation conditions that are *sufficient*, *atavistic*, and *informationally encapsulated*.

Consider disgust, an emotional and behavioral response that seems adapted, in the first case, for reacting to the threat of contamination (see Kelly 2011). It is a well-known feature of research into disgust that disgust can be activated by items

that pose no actual risk of contamination – for example, a sterilized cockroach, or fudge in the shape of dog poop. And this disgust can change subsequent behavior; for example, causing a chocolate lover to decline poop-shaped fudge (Rozin, Millman, and Nemeroff 1986). These responses exhibit the features mentioned above: merely looking like dog feces is enough to activate the response (sufficiency), the aversive response to such contamination seems a product of the natural history of humanity (atavism), and the background knowledge that the dog-poop-shaped candy is actually a delicious morsel does not prevent the activation from occurring (informational encapsulation). Disgust, then, is a perfect illustration that emotion can activate and alter behavior in cases that are out of sync with our all-things-considered judgments.[11]

While system 1 responses are sometimes activated by rather inflexible elicitors, it would be a mistake to assume that they are perfectly rigid. As critics of Haidt and Greene have noted, the activation of our real moral responses are often closely shaped by experience, circumstances, and background knowledge (Pizarro and Bloom 2003; Fine 2006; Narvaez 2010). Activation of automatic moral responses is frequently driven by processes that are not informationally encapsulated, and involve a considerable range of information to construe the situation (cf. Pizarro and Bloom 2003). The point is easy to see by considering an old philosophical example:

> Imagine a knock on your door. When you go to see who is there, you see your neighbor walking away in the distance, having left a brown envelope on your doorstep. When you open it, you find a *photograph* of your spouse having sex with your neighbor![12]

If you are like many people, you would nearly instantly experience a range of automatic emotional responses: anger, jealousy, and so forth. And you would feel them both toward the neighbor and toward your spouse. But now imagine a case just like this where instead of a photograph, you find that the envelope contains:

> A *drawing* of your spouse with the neighbor!

Now, your response might be much different. Among other things, your outrage might be directed at a different set of targets: still the neighbor, but not your spouse. The crucial point is that activation of your automatic emotional response is sensitive to your background knowledge – including knowledge about, say, the properties of photographs and drawings – with which you construe the situation (cf. Fodor 2000 on "the input problem"). Of course, this is consistent with emotions sometimes being triggered by appraisals that *are* sufficient and informationally encapsulated: the point is that activation of automatic processes can occur via multiple routes, some of which employ background knowledge, and others of which ignore it.

Additionally, moral responses may become automatic as the result of past intentional behavior (itself perhaps the product of reasoning and reflection). Conscious goal selection can result in altering automatic responses. Brandon Stewart and Keith Payne (2008) have shown this to be true in contexts of implicit racial attitude activation when participants make specific intentions to think counter-stereotypical thoughts upon encountering a black individual. Such subjects reduce their bias, but what is more, Stewart and Payne argue that this reduction proceeds by modification of *automatically produced* behavior, and with little practice.

The upshot is that while it is surely right to emphasize that our intuitive responses *are sometimes* inflexible, it is far from clear whether these inflexibilities typically lead us astray to the extent required if the literature on automaticity were to countenance a more general skepticism about moral judgment.

Are Intuitions Dominant?

We have just argued that even on the assumption that automatic processes are dominant in the production of our moral responses, this does not lead to skepticism without the additional – and highly debatable – assertion that automatic processes are unresponsive to the moral domain. We will now try blocking the road to skepticism with another tactic: questioning the dominance hypothesis itself.

To start, consider a range of ordinary situations:

- You are standing in the elevator and would like to hurry on to your floor. Through the door you see a person rushing towards the elevator. Should you push the "open door" button?
- You are driving on a backed-up freeway, and some drivers start to drive on the shoulder, bypassing stalled traffic. When the shoulder ends, these drivers attempt to merge back into the stream of traffic they bypassed. Should you let them in?
- After spending hours preparing a delicious vegetable stew using chicken stock, you recall that the acquaintance coming to dinner is a vegetarian. Would it be okay to not mention the stock and serve the stew to your acquaintance?
- An old but not especially close friend has mailed you a birthday gift for the last two years. Are you morally required to send them one in return?
- There is a small basket full of candy by the coffeemaker at work. How many pieces is it permissible to take?

We take it that these cases and numerous other everyday situations pose mundane moral questions, and that people respond to them both with judgments and behaviors. They are of course, a motley assortment: one might respond quite quickly to the time-sensitive elevator and traffic cases, dither for some time about

the stock and the birthday, and not give a thought to the candy dish (unless one came upon a colleague gobbling Hershey's Kisses by the score). In fact, the motleyness makes the point: if psychological intuitionism is correct, then typical moral responses to most or all everyday moral questions are dominated by the operation of system 1 (automatic) processes rather than system 2 (reasoning) processes, but it is not obvious that this is the case.

Mallon and Nichols (2011) point out that the dominance hypothesis assumes an answer to *the counting problem*. Since claims about the dominance of automatic processes are, in effect, claims about the proportion of real moral judgments or other behaviors that are driven by system 1 "intuitive" processes as opposed to those driven by system 2 "reasoning" processes, confirming these claims requires actually counting moral responses "in the wild" and assigning a causal explanation for them. However, to our knowledge no theorists have ever done such a thing, nor has anyone ever offered a methodology for determining which of the range of psychological causes of moral judgment identified in the laboratory are operating in typical moral judgments in everyday life. In fact, we know of no way of actually counting and classifying actual moral judgments to see whether these claims are true or false. Consider the ordinary situations we started with. How could you begin to count the moral judgments and behaviors in them, and to decide what cognitive processes were operative?

The counting problem also afflicts critics of psychological intuitionism. For example, Pizarro and Bloom suggest that Haidt underestimates the importance of reasoning processes in moral judgment, since reasoning can influence how one appraises a situation as well as controlling the inputs that shape one's automatic processes, but they *allow* that Haidt may be right that intuitive processes dominate in typical situations (Pizarro and Bloom 2003: 195, fn. 2). And Darcia Narvaez (2010) argues for the "partnership" of automatic and reasoning processes, but this claim, like the claim of dominance for automatic processes, faces the counting problem. In the absence of a solution to the counting problem, claims of the dominance of automatic processes and the denial of these claims look poorly supported by the available evidence.

Perhaps the psychological intuitionist should simply decline to give an answer to the counting problem. Counting, she might say, is a mug's game: knowing that a certain pathogen is implicated in a disease does not necessarily enable us to specify exactly how many cases have this etiology. Rather, her argument might be construed as a sort of inductive suspicion: given the ease with which the causal sufficiency of system 1 processes can be shown in the lab, the best guess is that they are widely dispersed in natural contexts – widely enough to make the dominance hypothesis the best supported generalization about moral judgment. Even supposing that there is no trouble with regards to "ecological validity" here, and the laboratory phenomena are readily observed in naturalistic contexts, there is still a problem about the induction base: very little of human life takes place in the lab, and that which does not is of a wildly heterogeneous nature (as indeed the examples just listed suggest).

The psychological intuitionist's inductive generalization would be stronger if it were buttressed with an argument explaining why the dominance of system 1 processes is to be expected. It turns out that there is such an argument, one Mallon and Nichols (2011) call *the finite resources argument*. On a dual process framework, reasoning (understood as a form of cognition) is a slow, effortful process that commands resources like attention that are themselves finite. Because of these resource limits, the thought goes that it is simply not possible for reasoning to subserve most of what we do. John Bargh and Tanya Chartrand put the point thus: "It may be hard to bear that most of daily life is driven by automatic, nonconscious mental processes – but it appears impossible . . . that conscious control could be up to the job" (1999: 464).

There are now numerous studies exploring the finite nature of volitional resources (Baumeister and Tierney 2011; Sripada 2010). For example, in one of Baumeister and colleagues' (1998: 1254–6) demonstrations of "ego depletion," participants were seated at a table bearing chocolate chip cookies, chocolate candies, and radishes. One group of participants was asked to eat only cookies or candies, while the other was asked to eat only radishes. Participants were then left alone for a few minutes to enjoy (or not enjoy) their snack. Although radish eating participants reported on the difficulty of resisting the chocolate, none actually broke their diet (the majority of chocolate eating participants, one presumes, had no comparable difficulty). Subsequently, participants were asked to solve two figure-tracing puzzles that were, unbeknownst to them, impossible to complete; the question was how long the participants would persist in this unrewarding task. The chocolate gobblers, it turns out, persisted more than twice as long as the radish nibblers (mean 18.90 minutes vs 8.35 minutes); the explanation is supposed to be that resisting temptation produces a "psychic cost," which left participants with less of whatever volitional wherewithal – "willpower," as Baumeister has it (Baumeister and Tierney 2011: 22–3) – is required for perseverance with difficult problems (Baumeister *et al.* 1998: 1255).

Does this sort of depletion translate into behavioral differences in real situations? Some evidence suggests so. In a remarkable recent study of Israeli judges, Shai Danziger and his colleagues (2011) found that the favorability of judges' verdicts to defendants starts off high following breaks, but then declines precipitously, recovering again after subsequent breaks. While there are many ways of interpreting these data, research on ego depletion makes it very tempting to understand the effect as the result of depletion of finite volitional reserves that are then replenished during the break by resting or consuming a snack. If so, we have got a case where finite resources look to play a role in influencing judgment and behavior, and provide an empirically grounded motivation for asserting the dominance of system 1 processes over system 2.

However, considerations from finite resources are only persuasive on the assumption that reasoning processes are *always* resource intensive. But this, Mallon and Nichols (2011) argue, need not be so. Mallon and Nichols propose a specific model for thinking otherwise, one on which a subject's reasoning employs a rule,

thereby operating without depleting finite resources of attention and willpower. Consider the following cases:

- A child is setting the table for a Thanksgiving meal, and is placing the salad forks on the right side of the plate.
- A sign on the highway reads:
  baltimore, next right.
- A driver changes lanes and then turns without using her turn signal.
- You attend a funeral service. Another attendee is wearing a cherry red suit.

In each case, if you are like us, you instantly apprehend that something is amiss, but you also find yourself ready and able to say what: The forks go on the left. "Baltimore" should be capitalized. Lateral moves should be signaled. Dark clothing is appropriate for a funeral.

Mallon and Nichols suggest that moral judgment (as well as other sorts of judgment) can be driven by moral rules that are deployed quickly and effortlessly in reasoning to resolve moral dilemmas. System 2 reasoning might simply take the form:

> Incest is wrong.
> That is a case of incest.
> That is wrong.

In this way, the depletion of finite resources can be bypassed.

There are other reasons to think this is plausible. First, in other work, Nichols and Mallon (2006) have offered independent reason for suggesting that moral rules may be operating in processing moral judgment: namely, that such rules could explain the different judgments seen in switch- and footbridge-style trolley cases (cf. Mallon and Nichols 2010). The idea, according to Nichols and Mallon, is that subjects endorse (or in some unconscious way implement) a moral rule that prohibits the second type of action (pushing the man off a footbridge), but not the first (redirecting the train).

Second, the existence, transmission, and enforcement of norms is surely one of the striking features of humanity (Sripada and Stich 2006). Indeed, the human capacity for norm-governed behavior likely played an important role in human evolution. Philip Kitcher, for example, has recently suggested that the capacity to follow rules can support behavioral altruism with a broader class of cooperators, allowing harmony in larger and larger social groups (Kitcher 2011).

Third, as gene-culture coevolutionary theories emphasize, culture is important to human evolution because it allows the transmission of information that adapts people to varied (and sometimes hostile) environments. But such power depends, in part, on the capacity of humans to employ it in ways that sometimes circumvent or override automated, prepotent responses (Richerson and Boyd 2005: 12).

Fourth, recent work on behavioral interventions (whose success depends on sparing scarce cognitive resources) has found that consciously endorsed

"implementation intentions" enhance control and reduce the effects of "ego depletion" (e.g., Stewart and Payne 2008; Webb and Sheeran 2003). Implementation intentions are intentions of the form, "As soon as situation y occurs, I will initiate goal-directed behavior x." They specify *how* one will act to carry out a goal. Webb and Sheeran (2003) found that subjects who endorsed implementation intentions regarding a Stroop task (wherein one has to say the color of a word which is itself the name of a different color) resisted ego depletion in performing the task, and were able to perform the task better when already ego depleted. Crucially, "by specifying when, where, and how one will act, implementation intentions pass control of behavior to anticipated environmental cues. Passing control to specified cues means that the need for cognitive control is circumvented" (Webb and Sherran 2003: 280). What Mallon and Nichols (2011) have suggested is, in effect, that moral rules can act like implementation intentions. They tell you: "Incest is wrong. Think no further. Proceed to judgment!" – a process that can use intentional level representations without exhausting precious resources.

But does not the Israeli Judge evidence count against this possibility? After all, the judges are presumably reasoning with explicit rules if anyone is. We concede that the study suggests that finite resources are in play. However, it is not obvious that it supports a dual process account in which system 1 processes dominate. First, it is not clear why system 1 processes should be systematically less favorable, and system 2 processes systematically more favorable, to defendants. But, more importantly, even if we interpret the drop off in favorability as the product of a shift from system 2 (resource intensive) mechanisms to system 1 (automatic) mechanisms, it remains far from clear why we should interpret the study as showing that system 1 processes rather than system 2 processes dominate. After all, when other resources are plentiful, system 2 processes seem up to the task of controlling behaviors in ways that differ systematically from those that would result from system 1.

A moral rules model is not a complete story of the mechanisms of moral judgment, but the existence of this model demonstrates the possibility of models of moral judgment that emphasize reasoning capacities and do not run afoul of the finite resources problem. This possibility, in turn, casts doubt on the finite resources problem as an argument for the dominance of automatic processes in ordinary moral judgment and behavior, suggesting that the counting problem remains a serious challenge to the dominance hypothesis.

Epiphenomenal Moral Reasoning

Defenders of the dominance hypothesis have one more line of attack, which suggests that "moral reasoning" is typically not the effective cause of behavior, but rather a mechanism for confabulating a post hoc justification for an antecedently

arrived at conclusion (Haidt 2001; Greene 2007). Recall that "moral reasoning" here can be understood in two ways. On the first, it is the operation of system 2 reasoning processes on problems in the moral domain. On the second, it is a sort of behavior, which could be undertaken in concert with other people to arrive at answers regarding problems in the moral domain, but which is also subserved in substantial part by system 2 reasoning. Haidt and Greene allow that moral reasoning in both senses sometimes plays a role in the production of moral judgment and behavior, but in endorsing the dominance hypothesis they hold that most of the time moral reasoning (in either sense) merely acts to rationalize the outcome of system 1 processes. In other words, they insist that moral reasoning is typically epiphenomenal with regard to the production of moral judgment.

Canonical evidence for such epiphenomenality of moral reasoning comes from cases of moral dumbfounding like the Julie and Mark case earlier. In these cases, it seems that subjects make a judgment, but are unable to justify the judgment, giving weight to the sense that their judgment was not the product of conscious reasoning. While Haidt's evidence for moral dumbfounding is sometimes unsystematic (cf. Wheatley and Haidt 2005), there is much evidence from a range of research programs that suggest people glibly confabulate explanations for their behaviors in some contexts (e.g., Hirstein 2005; Schnider 2008; Doris forthcoming).

To our knowledge, the most systematic exploration of the epiphenomenality of reasoning in the context of moral judgment has been the important web-based studies conducted by Fiery Cushman and colleagues. Cushman, Young, and Hauser (2006) elicited moral judgments and justifications for those judgments and coded them for their sufficiency, and found subjects sometimes unable to justify their own responses. For example, Cushman and colleagues used "trolley"-style cases to explore whether subjects upheld three moral principles: (1) that harm resulting from action is worse than that resulting from omission, (2) that intended harm is worse than accidental harm, and (3) that harm that results from bodily contact is worse than that which does not result from bodily contact. They found that subjects' responses were consistent with each principle. When subjects were asked to justify their responses to the vignettes, however, they seemed to have at best imperfect understanding of the principle involved. In action–omission cases and bodily contact cases, most subjects were able to offer a sufficient justification, while in intentional vs accidental harm cases, most were not. Moreover, in these intentional vs accidental cases, nearly a third of subjects' justifications either invoked features that were not actually present in the scenario (e.g., said a switch was already flipped when it was not) or claimed to have made an error in the original choice (i.e., "I pushed the wrong button"). Crucially, the authors take these to be a mark that the justification was a confabulation generated at the moment of justification.

We note first that the evidence from these studies does not obviously favor the dominance of system 1 processes or the epiphenomenality of reasons. As Cushman and colleagues note, subjects often *do* justify their judgments by appeal

to principles that would *both* explain and justify their responses. For this reason, Cushman and colleagues suggest that their data support the conclusion that multiple systems are involved in producing moral judgment, with the possibility that different systems operate for different principles.

Despite this ecumenical conclusion, it may seem that psychological intuitionists are still winning, for though subjects in Cushman and colleagues' studies sometimes were able to produce justifications for their judgments, they sometimes were not. What is more, in those cases where they did produce such justifications, they may still have arrived at them post facto, rather than by accessing the reasons that lead to their judgments. So psychological intuitionists can offer a deflationary explanation of the data.

However, just as intuitionists have an explanation of data that apparently favors reasoning processes, so defenders of a role for reasoning have a reply that undermines data in favor of intuition. Mallon and Nichols (2011) note a problem that infects both the anecdotal and experimental exploration of moral dumbfounding: the conflation of justification and explanation. Psychological intuitionists (and many of their critics) tend to assume that:

> If a judgment results from a reasoning process, it will be the product of mental states that not only *explain* it, but also *justify* it.

We can see that this assumption could be false by considering again the possibility that such judgments are driven by moral rules. Recall the possibility that subjects embrace a moral rule that "incest is wrong," and then, when faced with the case of Julie and Mark's incest, they simply reason: "Incest is wrong. And that is incest. So that is wrong." That bit of valid reasoning is sufficient to proximally *explain* a subject's disapproving judgment, but it fails to *justify* the conclusion if subjects cannot subsequently offer a justification of the premises. Findings of dumbfounding show that many subjects are not skilled at offering such justifications. But this shows nothing about the kind of cognitive process involved. It is, in fact, especially uncharitable to interpret a defender of the *causal efficacy* of reasoning processes as also being committed to the view that every reason itself is supported by other reasons that are themselves (1) coherent with other held principles, (2) consciously accessible, and (3) justified.

Things are more complicated if we consider how this objection applies to experimental work in this area of the sort pursued by Cushman and colleagues. Such work begins by offering subjects a series of vignettes to elicit moral judgments. Subjects are subsequently offered a pair of similar scenarios about which they have previously offered divergent judgments and asked to justify divergent responses. For example, a subject who judged killing one to save five permissible in a case like Switch, but judged killing one to save five impermissible in a case like Footbridge, could be asked to offer a justification for the difference.

Like the dumbfounding reported for the Julie and Mark experiment, this methodology is driven by the assumption that a failure of justification is a failure of

explanation. To be fair, Cushman and colleagues are sometimes sensitive to the distinction between explanation and justification. For instance, they allow that if a subject identifies a sufficient factual basis for a distinction but also disavows it as morally relevant, it should be coded as both "sufficient" (as an explanation) and "denial" (as a justification). On the other hand, they code as "failed" justifications instances where "conflicting principles applied separately to each case, e.g. 'In Chris's case it is wrong to murder, but in Candy's case you have to save the most people'." Since in the latter case subjects might be providing real explanations for their judgments that fail to be adequate justifications (because they lack holistic coherence), this way of coding conflates explaining with justifying.[13] The result is that it is somewhat difficult to interpret the experimental data from Cushman and colleagues in light of the distinction between explanation and justification.

The crucial cases are those where subjects make a judgment best explained by a certain principle, but where the subjects fail to appeal to that principle in explaining or justifying their judgment. Even in these cases, however, the failure does not show that *reasoning did not produce subjects' moral judgments* so much as that *subjects are bad at generating or reporting reasoning in support of such judgments.* Again, those defending the causal efficacy of reasoning processes need not maintain that such processes always operate on mental states that are themselves coherent with other held principles, consciously accessible, and justified.

We do not deny that intuitionist-style dumbfounding never occurs, but only that (1) the extant evidence and theory do not compel the conclusion that it always (or even usually) occurs and (2) the failure to distinguish explanation from justification considerably clouds the evidence for the epiphenomenality of reasoning. As we have suggested, we think it is plausible that at least some moral judgments result from moral rules which are themselves applied in reasoning to specific cases. We also think it plausible that reasoning can play a role both in activating automatic processes and in automating goals. The available data on failures of justification do not conclusively undermine these thoughts.

Culture and Moral Life

So far, we have considered how attention to automatic processes has, in the hands of psychological intuitionists, engendered skepticism about moral judgment. We have denied that this inference is mandatory. In closing, we will offer a diagnosis of the reasoning that might lead one to think it is mandatory. We contend that the culprit is a dubious assumption to the effect that human morality is dominated by cognitive processes whose shape is sensitive more to evolutionary history than to people's life history and circumstances. We see this assumption in the insistence of the dominance of system 1 processes, in the emphasis on their inflexibility, and in the suggestion that moral reasoning itself is epiphenomenal.

The assumption seems especially compelling when viewed in the long shadow cast by recent work in evolutionary psychology. Evolutionary psychology, paired with dual-process accounts of cognition, has identified system 1 processes with evolutionarily older, "short-leash" mechanisms (e.g., Stanovich 2004) conceived of as inflexible "alarm bells" (Greene 2007) shaped primarily by our evolutionary past. But what this way of thinking has left out of the picture is the extent to which individual experience and transmitted culture – understood as a set of theories, norms, and other information passed from individual to individual and population to population – shape moral thought about behavior.

We can illustrate our point by considering recent work in the evolutionary psychological tradition on incest avoidance. Debra Lieberman and colleagues (2003) have produced compelling evidence in favor of Westermark's hypothesis that childhood coresidence with an opposite-sex sibling predicts moral disapproval to incest among *third parties*. The effect is that children who did not experience such childhood coresidence have weaker moral sentiments against third-party incest. Lieberman, Tooby, and Cosmides thus suggest an explanation for our endorsement of cultural norms concerning incest that are rooted in the interaction of our endogenously determined cognitive architecture with the experience of psychological aversion to sex with one's coresidential siblings but which is then transferred to culturally transmitted, public representations of taboo acts.[14]

Lieberman and her colleagues use their data to defend the existence of a "human neural architecture [that] includes a specialized kin-recognition system that evolved among our hunter-gatherer ancestors [in part] . . . to inhibit sex among reproductively mature close genetic relatives because children produced from such unions would be less healthy" (2003: 820), and they contrast their view with one that would explain incest avoidance entirely as a product of inherited culture. That seems right as far as it goes but, by framing the story within a Manichean biology-vs-culture framework, they miss the opportunity to highlight the complex developmental story they really favor: one that has an important role both for individual experience and for inherited culture in determining one's level of support for moral norms concerning third-party incest.

The resulting shape of the public, cultural norm could then depend on three features: evolutionarily shaped cognitive mechanisms, individual developmental histories, and the specific, culturally transmitted content of moral rules. This fits nicely with an influential model on which the content of transmitted culture is fixed by biological constraints that make certain contents "good to think" (Sperber 1996). But do the contents of transmitted culture – for example, the content of moral norms – have any independent role to play in influencing moral behavior?

In fact, the view that transmitted culture plays *no* substantial role in fixing moral judgment and behavior is enormously implausible, and we know of no one who argues for it. Jesse Prinz (2007) has noted sexual norms restricting sexual activity among relatives exhibit cultural variation:

> In almost all cultures . . . there are prohibitions against incest, but these vary considerably. In Euro-American cultures it is considered incestuous to marry a first cousin, while in parts of India and Pakistan and the Middle East, marrying a first cousin is strongly encouraged . . . In some cultures incest between siblings is punished severely, and in others it is merely discouraged.
>
> (188–9)

Prinz draws attention to two sorts of diversity: diversity in what counts as incest, and diversity in the normatively right response to incest. The former speaks to differences in third-party incest norms that speak directly to the ability of cognitive mechanisms to exhaustively explain the content of those norms. The latter speaks to the fact that norms governing moral responses are underdetermined by those mechanisms.

More generally, there is widespread cultural variation in moral norms, something even defenders of moral nativism rarely deny, but which a great deal of the existing psychological research elides, since it only infrequently involves cross-cultural data (Henrich, Heine, and Norenzajan 2010). But where cross-cultural data are available, moral diversity is becoming increasingly evident. For example, a growing body of evidence from Joe Henrich and his colleagues (Henrich *et al.* 2005; Henrich *et al.* 2010) has documented cultural diversity in strategies pursued in standard economic games (in response to suggestions that these strategies are human universals). In the United States, Richard Nisbett and Dov Cohen (1996) have documented differences between Southern white males and their Northern counterparts along a range of behavioral and physiological measures – findings they attribute to the existence of a "culture of honor" among Southern white males. Here, as in the economic games cases, it is plausible that differences in judgment and behavior are driven by the cultural transmission of different norms.

Of course, as we have said, the obvious model to explain such phenomena is one on which moral judgments and behaviors are influenced by moral norms that emerge from extraordinarily intricate interactions among biological predispositions (usually assumed to be species-typical) ***and*** individual developmental experience ***and*** the content of culturally transmitted norms. Our view is that this obvious model is correct. This is not to say that some version of moral nativism is indefensible (for a defense, see Mikhail 2011). It is to say one cannot develop an adequate account of moral judgment and behavior without attention to the roles of individual experience and cultural transmission – a fact that we think will become increasingly apparent in the years to come but attention to which seems left out of the recent enthusiasm for psychological intuitionism.

Of course, variation in moral norms has long played a central role in arguments for moral skepticism, or "antirealism" (Mackie 1977; Loeb 1998; Doris and Stich 2005; Doris and Plakias 2007). But, if what we have said here is correct, the existence of evaluative diversity may play quite the opposite role in moral psychology, for it suggests that one important skeptical position, psychological intuitionism

about moral judgment, may be on quite uncertain footing. Evaluative diversity manifests the sensitivity of the psychological processes undergirding morality to cultural and individual circumstances – a sensitivity that the recent emphasis on "the automaticity of morals" has not yet fully acknowledged.

Although we have been urging caution regarding specific arguments in empirically informed ethics, we are sanguine about the long-term prospects of this research program. While the research program can already count many achievements, we expect and hope its most abiding contribution will be to replace seemingly intractable empirical debates within traditional, speculative approaches to philosophical ethics and moral psychology – over the role of emotions in the production of moral judgment and behavior, for example, or over the existence and character of moral agreement and disagreement across cultures – with more carefully specified and well-confirmed theories, rooted in the best traditions of the social sciences.

Notes

1 For overviews of the field, see Andreou (2007); Doris and Stich (2005, 2006). For collections containing representative work, see Doris and The Moral Psychology Research Group (2010); Knobe and Nichols (2008); Sinnott-Armstrong (2007a, b, c).

2 http://ns.umich.edu/htdocs/releases/story.php?id=8553 (accessed November 13, 2012).

3 For example, http://www.3ammagazine.com/3am/mind-reader/ (accessed November 13, 2012), http://www.nytimes.com/roomfordebate/2010/08/19/x-phis-new-take-on-old-problems/philosophy-vs-imitation-psychology (accessed November 13, 2012)

4 This could be, as Joshua Greene's work suggests (Greene *et al.* 2001), because system 1 processes dominate system 2 processes in a different way, namely, that when the two systems offer conflicting outcomes, system 1 processes typically win. We take no stand on this further issue here, except to note that such claims are afflicted by something analogous to the "counting problem" we discuss in the section "Debunking Arguments."

5 Greene has since backed away from the "impersonal/personal" distinction as a way of characterizing the distinction for reasons that, while compelling, make no difference to our present discussion. See Greene *et al.* (2008); McGuire *et al.* (2009); Greene (2009); Greene *et al.* (2009).

6 As a number of commentators have noted (e.g., Mason 2011), this way of casting the debunking argument seems to reflect a view on which at least some features of morality are objective – existing independently in a way that can lead us to either successfully apprehend them or fail to do so. While this view of the metaphysics of morality is contentious, for present purposes we simply concede this objectivism to the skeptic. This is because it may well be that the skeptical argument can be cast without it (as in Joyce 2006). And in any case, we aim to show that even if the objectivity of morality

is granted, skepticism looms not – or rather, that psychological intuitionism does not straightforwardly support its looming.

7 A second sort of argument reaches the same conclusion, but via a different route. Instead of arguing that our moral responses track some property or properties that are not morally relevant, this one suggests that our responses fail to satisfy one or more relatively formal constraints to serve as a source of knowledge. For example, if our moral responses are subject to order effects, framing effects, and the like then they may not exhibit the right sort of stability or coherence to be reliable or true (Horowitz 1998; Sinnott-Armstrong 2007d; Doris and Stich 2005; Sunstein 2005).

8 Greene has since modified this account in response to ongoing evidence and scrutiny (see n. 5).

9 These findings dovetail nicely with more recent work showing that patients with prefrontal cortex damage are more consequentialist (Koenigs *et al.* 2007), and that antisocial personality tendencies lead individuals to be more consequentialist (Bartels and Pizarro 2011). Of course, if some version of consequentialism is right, then processes that deviate from consequentialist reasoning count as distorting moral judgment, and these sorts of brain damage would enhance rather than impair moral judgment. (More generally, if we could determine which (if any) substantive moral theory is correct, we could more clearly judge which (if any) psychological processes reliably lead us to moral truth.)

10 We are grateful to Isaac Wiegman for discussion of this point.

11 And, as an increasing body of evidence also shows, disgust seems to play a prominent role in our moral lives (Haidt, Koller, and Dias 1993; Wheatley and Haidt 2005; cf. Kelly 2011 for review).

12 We heard of a case like this from Alan Leslie.

13 See: http://moral.wjh.harvard.edu/methods/resources/3prin_criteria.pdf (accessed November 13, 2012).

14 In fact, what they suggest could amount to a quite general mechanism for the genesis of many or all third-party moral rules, one on which they emerge and are endorsed via some sort of pairing with first-person feelings of aversion towards action types. This account is similar to Blair's (1995) model, the crucial feature of which is the transferral and abstraction of an aversion to moral violations in the first/second person to norms that concern third parties.

References

Andreou, C. (2007) "Morality and Psychology," *Philosophy Compass* 2: 46–55.

Bargh, J.A. and Chartrand, T.L. (1999) "The Unbearable Automaticity of Being," *American Psychologist* 54: 462–79.

Bartels, Daniel and Pizarro, David (2011) "The Mismeasure of Morals: Antisocial Personality Traits Predict Utilitarian Responses to Moral Dilemmas," *Cognition* 121: 154–61

Baumeister, R.E., Bratslavsky, E., Muraven, M., and Tice, D.M. (1998) "Ego Depletion: Is The Active Self a Limited Resource?" *Journal of Personality and Social Psychology* 74: 1252–65.

Baumeister, R.F. and Tierney, J. (2011) *Willpower: Rediscovering the Greatest Human Strength*, New York: Penguin.

Blair, R.J. (1995) "A Cognitive Developmental Approach to Morality: Investigating the Psychopath," *Cognition* 57 (1): 1–29.

Blair, R.J. (1996) "Brief Report: Morality and the Autistic Child," *Journal of Autism and Developmental Disorders* 26 (5): 571–9.

Blair, R.J., Mitchell, D.R., and Blair, K. (2005) *The Psychopath: Emotion and the Brain*, Oxford: Blackwell.

Chaiken, S. and Trope, Y., eds. (1999) *Dual-Process Theories in Social Psychology*, New York and London: The Guilford Press.

Cushman, F., Young, L., and Hauser, M. (2006) "The Role of Conscious Reasoning and Intuitions in Moral Judgment: Testing Three Principles of Harm," *Psychological Science* 17 (12): 1082–9.

Damasio, A.R. (1994) *Descartes' Error: Emotion, Reason, and the Human Brain*, New York: G.P. Putnam's Sons.

Danziger, S., Levav, J., and Avnaim-Pesso, L. (2011) "Extraneous Factors in Judicial Decisions," *PNAS*, April 26, 108 (17): 6889–92.

Darley, J.M. and Schultz, T.R. (1990) "Moral Rules: Their Content and Acquisition," *Annual Review of Psychology* 41: 525–56.

Doris, J.M. (2002) *Lack of Character: Personality and Moral Behavior*, Cambridge University Press.

Doris, J.M. (2005) "Replies: Evidence and Sensibility," *Philosophy and Phenomenological Research* 71 (3): 656–77.

Doris, J.M. (forthcoming) *Talking to Our Selves: Reflection, Skepticism, and Agency*, Oxford: Oxford University Press.

Doris, J.M. and The Moral Psychology Research Group, eds. (2010) *The Moral Psychology Handbook*, Oxford: Oxford University Press.

Doris, J.M. and Plakias, A. (2007) "How to Argue About Disagreement: Evaluative Diversity and Moral Realism," in *Moral Psychology, Volume 2: The Cognitive Science of Morality: Intuition and Diversity*, ed. W. Sinnott-Armstrong, Oxford: Oxford University Press, pp. 303–32.

Doris, J.M. and Stich, S.P. (2005) "As a Matter of Fact: Empirical Perspectives on Ethics," in *The Oxford Handbook of Contemporary Philosophy*, eds. F. Jackson and M. Smith, Oxford: Oxford University Press, pp. 114–52.

Doris, J.M. and Stich, S.P. (2006) "Moral Psychology: Empirical Approaches," in *The Stanford Encyclopedia of Philosophy*, Winter 2003 edn, ed. Edward N. Zalta, http://plato.stanford.edu/entries/moral-psych-emp/ (accessed November 13, 2012).

Evans, J.S.B.T. and Frankish, K. (2009) *In Two Minds: Dual Processes and Beyond*, Oxford: Oxford University Press.

Fine, C. (2006) "Is the Emotional Dog Wagging the Rational Tail or Chasing It? Unleashing Reason in Haidt's Social Intuitionist Model of Moral Judgment," *Philosophical Explorations* 9 (1): 83–98.

Fodor, J.A. (2000) *The Mind Doesn't Work That Way: The Scope and Limits of Computational Psychology*, Cambridge, MA: MIT Press.

Foot, P. (1978) "The Problem of Abortion and the Doctrine of the Double Effect," in *Virtues and Vices*, Oxford: Basil Blackwell, pp. 19–32.

Greene, J.D. (2003) "From Neural 'Is' to Moral 'Ought': What Are the Moral Implications of Neuroscientific Moral Psychology?" *Nature Reviews Neuroscience* 4: 847–50.

Greene, J.D. (2007) "The Secret Joke of Kant's Sou," in *Moral Psychology, Volume 3, The Neuroscience of Morality: Emotion, Brain Disorders, and Development*, ed. W. Sinnott-Armstrong, Cambridge, MA: MIT Press, pp. 35–117.

Greene, J.D. (2009) "Dual-Process Morality and the Personal/Impersonal Distinction: A Reply to McGuire, Langdon, Coltheart, and Mackenzie," *Journal of Experimental Social Psychology* 45 (3): 581–4.

Greene, J.D. (forthcoming) *For the Greater Good: How the Moral Brain Works and How It Can Work Better*, New York: Penguin Press.

Greene, J.D., Cushman, F.A., Stewart, L.E. *et al.* (2009) "Pushing Moral Buttons: The Interaction between Personal Force and Intention in Moral Judgment," *Cognition* 111 (3): 364–71.

Greene, J.D. and Haidt, J. (2002) "How (and Where) Does Moral Judgment Work?" *Trends in Cognitive Sciences* 6 (12): 517–23.

Greene, J.D., Morelli, S.A., Lowenberg, K. *et al.* (2008) "Cognitive Load Selectively Interferes With Utilitarian Moral Judgment," *Cognition* 107: 1144–54.

Greene, J.D., Sommerville, R.B., Nystrom, L.E. *et al.* (2001) "An fMRI Investigation of Emotional Engagement in Moral Judgment," *Science* 293: 2105–8.

Haidt, J. (2001) "The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment," *Psychological Review* 108: 814–34.

Haidt, J., Bjorklund, F., and Murphy, S. (n.d.) Moral Dumbfounding: When Intuition Finds No Reason, Unpublished manuscript, University of Virginia.

Haidt, J., Koller, S.H., and Dias, M.G. (1993) "Affect, Culture, and Morality, or Is It Wrong to Eat Your Dog?" *Journal of Personality and Social Psychology* 65: 613–28.

Henrich, J., Boyd, R., Bowles, S. *et al.* (2005) "'Economic Man' in Cross-Cultural Perspective: Behavioral Experiments in 15 Small-Scale Societies," *Behavioral and Brain Sciences* 28 (6): 795–815.

Henrich, J., Ensminger, J., McElreath, R. *et al.* (2010) "Markets, Religion, Community Size, and the Evolution of Fairness and Punishment," *Science* 327 (5972): 1480–4.

Henrich, J., Heine, S., and Norenzayan, A. (2010) "The Weirdest People in the World?" *Behavioral and Brain Sciences* 33 (2–3): 1–75.

Hirstein, W. (2005) *Brain Fiction: Self-Deception and the Riddle of Confabulation*, Cambridge, MA and London: MIT Press.

Horowitz, T. (1998) "Philosophical Intuitions and Psychological Theory," in *Rethinking Intuition: The Psychology of Intuition and its Role in Philosophical Inquiry*, eds. M. DePaul and W. Ramsey, Lanham, MD: Rowman & Littlefield, pp. 143–60.

Joyce, R. (2006) *The Evolution of Morality*, Life and Mind: Philisophical Issues in Biology and Psychology, Cambridge, MA: MIT Press.

Kelly, D. (2011) *Yuck! The Nature and Moral Significance of Disgust*, Cambridge, MA: MIT Press.

Kennett, Jeanette and Fine, Cordelia (2009) "Would the Real Moral Judgment Please Stand Up? The Implications of Social Intuitionist Models of Cognition for Meta-ethics and Moral Psychology," *Ethical Theory and Moral Practice* 12: 77–96.

Kitcher, P. (2011) *The Ethical Project*, Cambridge, MA: Harvard University Press.

Koenigs, M., Young, L., Adolphs, R. *et al.* (2007) "Damage to the Prefrontal Cortext Increases Utilitarian Moral Judgments," *Nature* 446: 908–11.

Kohlberg, L. (1969) "Stage and Sequence: The Cognitive-Developmental Approach to Socialization," in *Handbook of Socialization Theory and Research*, ed. D. Goslin, Chicago: Rand McNally, pp. 347–480.

Knobe, J. and Nichols, S. (2008) *Experimental Philosophy*, Oxford: Oxford University Press.

Lazarus, R. (1994) "Universal Antecedents of the Emotions," *The Nature of Emotion: Fundamental Questions*, eds. P. Ekman and R. Davidson, New York: Oxford University Press, pp. 163–71.

Leslie, A.M., Mallon, R., and DiCorcia, J.A. (2006) "Transgressors, Victims, and Cry Babies: Is Basic Moral Judgment Spared in Autism?" *Social Neuroscience* 1 (3–4): 270–83.

Lieberman, D., Tooby, J., and Cosmides, L. (2003) "Does Morality Have a Biological Basis? An Empirical Test of the Factors Governing Moral Sentiments Regarding Incest," *Proceedings of the Royal Society B* 270: 819–26.

Loeb, D. (1998) "Moral Realism and the Argument from Disagreement," *Philosophical Studies* 90: 281–303.

Mackie, J.L. (1977) *Ethics: Inventing Right and Wrong*, New York: Penguin.

Maibom, Heidi (2005) "Moral Unreason: The Case of Psychopathy," *Mind & Language* 20: 237–57.

Maibom, Heidi (2010) "What Experimental Evidence Shows Us about the Role of Emotions in Moral Judgment," *Philosophy Compass* 5 (11): 999–1012.

Mallon, R. and Nichols, S. (2010) "Rules," in *The Oxford Handbook of Moral Psychology*, eds. J. Doris and The Moral Psychology Research Group, Oxford: Oxford University Press, pp. 297–320.

Mallon, R. and Nichols, S. (2011) "Dual Processes and Moral Rules," *Emotions Review* 3: 284–6.

Mason, Kelby (2011) "Moral Psychology and Moral Intuition: A Pox on All Your Houses," *Australasian Journal of Philosophy* 89 (3): 441–58.

McGuire, J., Langdon, R., Coltheart, M., and Mackenzie, C. (2009) "A Reanalysis of the Personal/Impersonal Distinction in Moral Psychology Research," *Journal of Experimental Social Psychology* 45 (3): 577–80.

Mikhail, J.M. (2000) "Rawls' Linguistic Analogy: A Study of the 'Generative Grammar' Model of Moral Theory Described by John Rawls in 'A Theory of Justice'," PhD Thesis, Cornell University.

Mikhail, J.M. (2011) *Elements of Moral Cognition: Rawls' Linguistic Analogy and the Cognitive Science of Moral and Legal Judgment*, New York: Cambridge University Press.

Narvaez, D. (2010) "Moral Complexity: The Fatal Attraction of Truthiness and the Importance of Mature Moral Functioning," *Perspectives on Psychological Science* 5 (2): 163–81.

Nichols, S. (2004) *Sentimental Rules: On the Natural Foundations of Moral Judgment*, Oxford and New York: Oxford University Press.

Nichols, S. and Mallon, R. (2006) "Moral Rules and Moral Dilemmas," *Cognition* 100: 530–42.

Nisbett, R.E. and Cohen, D. (1996) *Culture of Honor: The Psychology of Violence in the South*, Boulder, CO: Westview Press.

Nisbett, R.E. and Wilson, T. (1977) "Telling More Than We Can Know: Verbal Reports on Mental Processes," *Psychological Review* 84: 231–59.

O'Neill, P. and Petrinovich, L. (1998) "A Preliminary Cross-Cultural Study of Moral Intuitions," *Evolution and Human Behavior* 19 (6): 349–67.

Petrinovich, L., O'Neill, P., and Jorgensen, M. (1993) "An Empirical Study of Moral Intuitions: Toward An Evolutionary Ethics," *Journal of Personality and Social Psychology* 64 (3): 467–78.

Pizarro, D.A. and Bloom, P. (2003) "The Intelligence of the Moral Intuitions: A Reply to Haidt (2001)," *Psychological Review* 110:193–6.

Prinz, J.J. (2007a) *The Emotional Construction of Morals*, Oxford: Oxford University Press.

Quinn, Warren (1989) "Actions, Intentions, and Consequences: The Doctrine of Double Effect," *Philosophy and Public Affairs* 18 (4): 334–51.

Richerson, Peter and Boyd, Robert (2005) *Not By Genes Alone*, Chicago: University of Chicago Press.

Rozin, P., Millman, L., and Nemeroff, C. (1986) "Operation of the Laws of Sympathetic Magic in Disgust and Other Domain," *Journal of Personality and Social Psychology* 50: 703–12.

Schnider, A. (2008) *The Confabulating Mind: How the Brain Creates Reality*, Oxford: Oxford University Press.

Singer, Peter (2005) "Ethics and Intuitions," *Journal of Ethics* 9 (3–4): 331–52.

Sinnott-Armstrong, W., ed. (2007a) *Moral Psychology, Volume 1: The Evolution of Morality: Adaptations and Innateness*, Cambridge, MA: MIT Press.

Sinnott-Armstrong, W., ed. (2007b) *Moral Psychology, Volume 2: The Cognitive Science of Morality: Intuition and Diversity*, Cambridge, MA: MIT Press.

Sinnott-Armstrong, W., ed. (2007c) *Moral Psychology, Volume 3: The Neuroscience of Morality: Emotion, Brain Disorders, and Development*, Cambridge, MA: MIT Press.

Sinnott-Armstrong, W., ed. (2007d) "Framing Moral Intuitions," *Moral Psychology, Volume 2: The Cognitive Science of Morality: Intuition and Diversity*, Cambridge, MA: MIT Press.

Sorensen, Roy (1991) " 'P, Therefore, P' without Circularity," *Journal of Philosophy* 88 (5) 245–66.

Sorensen, Roy (1996) "Unbeggable Questions," *Analysis* 56 (1): 51–5.

Sorensen, Roy (1999) "An Empathic Theory of Circularity," *Australasian Journal of Philosophy* 77 (4): 498–509.

Sperber, Dan (1996) *Explaining Culture: A Naturalistic Approach*, Oxford: Blackwell.

Sripada, Chandra Sekhar (2010) "*Philosophical Questions about the Nature of Willpower*," *Philosophy Compass* 5 (9): 793–805.

Sripada, Chandra Sekhar and Stich, S.P. (2006) "A Framework for the Psychology of Norms," in *The Innate Mind: Culture and Cognition*, eds. P. Carruthers, S. Laurence, and S.P. Stich, New York: Oxford University Press, pp. 280–301.

Stanovich, K.E. (2004) *The Robot's Rebellion: Finding Meaning in the Age of Darwin*, Chicago: University of Chicago Press.

Stewart, B.D. and Payne, B.K. (2008) "Bringing Automatic Stereotyping Under Control: Implementation Intentions as Efficient Means of Thought Control," *Personality and Social Psychology Bulletin* 34: 1332–45.

Stich, S.P. (1990) *The Fragmentation of Reason: A Preface to a Pragmatic Theory of Cognitive Evaluation*, Cambridge, MA: MIT Press, A Bradford Book.

Stich, S.P., Doris, J.M., and Roedder, E. (2010) "Altruism," in *The Moral Psychology Handbook*, eds. J.M. Doris and The Moral Psychology Research Group, Oxford: Oxford University Press, pp. 147–205.

Sunstein, C. (2005) "Moral Heuristics," *Behavioral and Brain Sciences* 28: 531–42.

Thomson, J. (1976) "Killing, Letting Die, and the Trolley Problem," *The Monist* 59: 204–17.

Webb, T. and Sheeran, P. (2003) "Can Implementation Intentions Help to Overcome Ego-Depletion?" *Journal of Experimental Social Psychology* 39: 279–86.

Wheatley, T. and Haidt, J. (2005) "Hypnotically Induced Disgust Makes Moral Judgments More Severe," *Psychological Science* 16: 780–4.

Wilson, T.D. (2002) *Strangers to Ourselves: Discovering the Adaptive Unconscious*, New York: Belknap.

Chapter 9

The Relevance of Responsibility to Morality

Ingmar Persson

Three Ways in Which Responsibility Is Relevant to Morality

Morality would be superfluous unless there were agents who were capable of being responsible for their acts or, in other words, who could be justifiably praised or blamed for having acted rightly or wrongly. But responsibility is a philosophically controversial notion. It is a familiar claim that it is incompatible with determinism, the doctrine that everything that happens in the world has a sufficient cause. So, an investigation into the nature of responsibility might show that there is no such thing as morality because determinism rules. Although I reject this radical conclusion, I shall argue that responsibility bears upon commonsense morality in more than one way and that a philosophical exploration of these diverse ways will show that commonsense morality needs to be revised. Let us begin by trying to map the ways in which responsibility bears upon commonsense morality.

As already remarked, there would not be any need for morality, as a set of norms or principles about what acts are morally right or wrong, unless there were agents capable of being responsible for acting morally rightly and wrongly. In particular, there must be agents who possess the mental equipment to act morally. Plausibly, this comprises that they have some understanding of what actions are morally right and wrong, and also about what it is that makes them morally right and wrong. In addition, they must have powers of action which enable them to act out their views about what actions are morally right and wrong. Otherwise, they could not be morally responsible for their actions; that is, they could not be justifiably blamed or praised, punished or rewarded, for their actions.

Praising somebody is expressing some positive attitude, for example liking, gratitude, or admiration, towards this individual, and blaming somebody is

The Blackwell Guide to Ethical Theory, Second Edition. Edited by Hugh LaFollette and Ingmar Persson.

expressing some negative attitude, for example dislike, anger, or contempt, towards them.[1] There is a wide spectrum of these attitudinal expressions. One form of expression is in one's private thoughts, but expression could also be behavioral and consist in various forms of treatment that benefit or harm, and, thus, constitute rewarding and punishing. Among these forms of treatment I would like to include our tendency to associate with the people whom we praise and dissociate from the people whom we blame since, for social creatures like us, to be the target of these forms of treatment is desirable and undesirable, respectively. By contrast, T.M. Scanlon rejects the view that blame "is a milder form of punishment" (2008: 122), but what he says about blame – that it is an attitude made appropriate by some impairment of social relations due to the behavior of the blamee – seems to me largely consistent with my view. In any case, I shall assume – to simplify exposition – that there is only a difference of degree between praising and rewarding, and blaming and punishing.

Agents need not be responsible on every occasion they act rightly or wrongly, for somebody who acts wrongly might be (partly or wholly) *excused* for doing so. Perhaps some agents had no way of knowing some relevant facts – for example, that the cargo they handled contained explosives so they could not have known that they ought to have handled it more gently, or they were so emotionally upset that they were unable to handle it sufficiently gently. Excuses have the function of removing or mitigating responsibility, and they are applicable only to agents who would otherwise be responsible, and not to inanimate things, plants, or animals that permanently lack the capacity for responsible actions.

But agents who possess the mental capacities to be morally responsible would not be able to act morally rightly and wrongly unless there were other individuals whom they could affect for better or worse. It seems plausible to hypothesize that you could act morally rightly and wrongly to someone only if something could be *good or bad* for them, that is, only if they are subjects of welfare or well-being. It also seems reasonable, though it is disputed by some, that things could be good or bad *for* individuals only if they have desires and feelings and, thus, consciousness. In order for something to be good or bad for someone, it appears that they would have to be capable of feeling it to be pleasant or unpleasant, of wanting it or being averse to it, and so on. In other words, a consciousness which encompasses feelings and desires seems necessary for possession of moral status or standing.

Some might wish to phrase the condition that possession of consciousness is necessary for moral status so weakly it allows that in order for it to be possible now to act morally rightly or wrongly towards a being, it need not have consciousness *now*; it is enough if it is endowed with a *potential* to develop consciousness in the future. If this is correct, you could treat a preconscious fetus wrongly by aborting it if you thereby prevent it from developing consciousness and leading a life that would be good for it.

Others claim that what endows an individual with moral status is not the fact that things could be good or bad for it, but that it has (moral) *rights*, for example,

to its life and limbs. But it seems that someone could be a rights holder only if things could be good or bad for them: we do not say that someone has a moral right to something X, unless we think that they could benefit from X. Rights to things that are of no use to rights holders would be pointless, and would be waived by the putative rights holders.[2] Thus, even according to a rights theory being a subject of well-being or welfare would be necessary for having moral status. However, on a rights theory the class of individuals who have moral status might be narrower, since the condition of being a subject of well-being might not be sufficient. Arguably, being a rights holder requires a more sophisticated mental set than being a subject of welfare does. For instance, things can plausibly be good or bad for a fly – because it can probably feel pain – but it seems wrongheaded to ascribe rights to it. Furthermore, we saw it might be that the consciousness condition should be weakened to require only a potential to develop consciousness. But it seems that organisms could not be right holders in virtue of having merely such a potential. Consequently, according to a theory of rights, the class of individuals in possession of moral status is likely to be smaller than it is on a welfare theory, but the requirement of being a subject of welfare would still retain its relevance.

Commonsense morality contains principles about what is a *just or fair* distribution of well-being to individuals with moral status. According to one of these principles, it is just that they get the degree of well-being that they *deserve*. If we acknowledge rights, we are likely to hold also that it is just that they have that to which they have rights. Another, more formal, principle of justice appears to be to the effect that when there is nothing, such as desert or rights, to make it just that some individuals have more welfare than others, an *equal* distribution of welfare is just. Attribution of desert is another point at which responsibility enters into morality, because it is reasonable to maintain that in order for someone to deserve to be rewarded or punished for what they have done, they must be responsible for it. Thus, responsibility is implicated not only in an agent's process of deliberation – because the agent must exercise psychological capacities requisite for responsibility – but also in the content of deliberation when it concerns the distribution of well-being to recipients.

There is a third context in which responsibility enters into morality, alongside the deliberative and the distributive contexts: it enters into the relations between agents and recipients in constraints on what agents are permitted to do to promote the well-being of recipients. According to the *act–omission doctrine*,[3] it is more difficult to justify doing harm of certain kinds than letting this sort of harm occur by omitting to act. For example, this doctrine prohibits killing one innocent (and nonconsenting) person in order to save the lives of five other innocent people, though the saving of five innocents is a greater good than the killing of one innocent is a harm. We have seen that a type of act's being morally wrong presupposes the existence of agents who could be responsible for the performance of this type of act. Hence, if an act, like killing an innocent, is harder to justify morally than letting an innocent die, responsibility for the outcome of an act is likely to be greater than responsibility for the same outcome of an omission. The value of the

outcomes being the same, it must be the different relations to them that explain the moral difference.

Now, it is possible to have a theory of rights according to which it is as wrong to let rights be violated by others as it is to violate them oneself.[4] According to such a theory, it would be permissible, indeed even obligatory, to kill one innocent if this was necessary to prevent five other innocents from having their rights violated by being killed. But this is not the kind of rights theory that commonsense morality endorses; it endorses a theory according to which it is much more wrong to cause right-violating deaths than to let such deaths occur – indeed, so much more wrong that it would take the saving of a very large number of lives to justify the causing of one right-violating death, if indeed this could be justified at all.

There is also a doctrine which could be construed as loosening the constraint on rights-infringing acts that the act–omission doctrine imposes. According to the act–omission doctrine, there could be situations in which no action is permissible because they all involve the killing of innocents. Suppose you are driving an unstoppable vehicle and that at a fork you have to turn either to the left or to the right, but if you turn left, you will kill one innocent person, and if you turn right, you will kill five innocent people. The *doctrine of the double effect* has been called upon to make it permissible for you to turn left in order to save the five because it says that it is permissible to cause smaller harm in the pursuit of a greater good if this is not intended as a means, or as an end, but is merely a foreseen side effect. In contrast, you are permitted to enable yourself to save five even by means of letting one die.

But what if your intention in turning left is to kill the one and you merely foresee that you will save the five? It seems that this cannot make your *action* of killing the one wrong, since it is natural to say that this is the same action as it would have been had you intended to save the five. To my mind, it is more natural to say that you have performed the right action for a *wrong reason*. Consequently, having the right intention is not necessary for acting rightly; it is, rather, necessary for not being *blameworthy*, or for not being a person who exhibits moral badness.

Moreover, a case by Frances Kamm (1996: 151) indicates that it is actually irrelevant to the wrongness of your act whether you intend a smaller harm in the pursuit of a greater good, or merely foresee it. Imagine that in order to save the five from being killed by an unstoppable vehicle, you will have to blow it up by throwing a grenade at it. Unfortunately, this will kill an innocent bystander as a side effect. Then, according to commonsense morality, you are not permitted to do this, even though you intend to save the five and merely foresee that the one will be killed. This example seems to indicate that it is crucial whether you *initiate a new cause of death*, or merely manipulate a preexistent threat, as in the case of the vehicle. The reason why this is relevant may be that in the latter sort of case your causal contribution to the death of the one is felt to be smaller, since whatever gave rise to the preexistent threat also makes a contribution. Analogously, if you save five by means of killing one, the causal link to the greater good is longer and, thereby, weaker than the causal link to the smaller harm of killing

the one. If this is right, causal considerations lie at the bottom of the doctrine of the double effect. As will emerge in the section "Responsibility and Causation," the same is true of the act–omission doctrine.

Accordingly, I shall call this third context, in which responsibility is relevant to morality, the causally differential context because the issue in this context is whether responsibility is a function of making a causal difference. Thus, the three contexts in which responsibility is relevant to morality are the deliberative, the distributive, and the (causally) differential contexts.

Responsibility and Reasons for Action

After this survey of three moral contexts of responsibility, let us first look closer at the context of deliberation. As remarked, being capable of performing morally responsible actions involves having a conception of what is morally right or wrong and being able to put this conception into effect in action. In other words, it involves the capacity to consider moral reasons and to act upon them. This suggests a view like Scanlon's that "'being responsible' is mainly a matter of the appropriateness of demanding reasons." Scanlon goes on to claim that because this is so, for there to be responsibility

> it is enough that the attitude in question be a judgment-sensitive one – that is, one that either directly reflects the agent's judgment or is supposed to be governed by it. For this reason, one can be responsible not only for one's actions but also for intentions, beliefs, and other attitudes.
>
> (1998: 22; cf. 272)

To be sure, actions are not "judgment-sensitive attitudes," but when they are intentional – and this is a paradigm case of being responsible for them – they are "the expression of judgment-sensitive attitudes" (Scanlon 1998: 21), for example, intentions. When we act for reasons for action, whose content is about what happens when we perform various actions, then, on the basis of consideration of those reasons, we form judgment-sensitive (propositional) attitudes, directed at those actions the reasons are about: we desire, decide, intend, and try to perform these actions. If in addition we have the requisite abilities and opportunities of action, then intentional actions will issue. It is a controversial issue, which is beyond the scope of this essay, as to exactly how an action must "issue" from desires, intentions, and so on, in order to be intentional.

It is true that it is not necessary that an act be intentional for the agent to be responsible for it: we are responsible for what we do out of negligence, though this is not intentionally done. However, in these situations some action by which we negligently risk causing something is still intentional, for example, we intentionally handle the goods whose explosiveness we neglect.[5]

The principal problem with Scanlon's account is not that it is too narrow, by excluding responsibility for intentional actions, but that it is too broad. There are judgment-sensitive attitudes for which we are not responsible. For instance, it seems that we are often not responsible for what we *believe*, even though we have (good) reasons for our beliefs. When I see the traffic light change from red to amber, I have good reason to believe that it will soon be green, and will immediately acquire this belief. But I can hardly be said to be responsible for having this belief; nor for my belief that the light is green when I see its greenness. Our beliefs can be *rational or irrational*, and can in this sense comply or fail to comply with norms, without it seeming right to hold us responsible for them. As these examples illustrate, this is so when we immediately form beliefs in response to sense experience, or reach them by immediate inference. It is less clear that we are not responsible for our beliefs when we form them as the outcome of deliberation upon reasons for and against, since we could be responsible for the deliberative process itself. But, for a reason that will surface soon, I deny that we are responsible for our beliefs even in such cases.

A criticism similar to the one here made of Scanlon can be made of Philip Pettit and Michael Smith when they claim: "To hold a belief or desire freely is to hold it in the presence of an ability, should the belief or desire be wrong, to get it right" (1996: 445).[6] Somebody has the latter ability if "he would come around to the right belief in the event of your pressing him with the demands of the evidence" (1996: 446). To my ears, such a sensitivity to evidence sounds more like the mark of rationality, of holding a belief or desire rationally. Pettit and Smith are conscious of the fact that, in thinking that beliefs are as much a matter of freedom and responsibility as the will, they go against "the prevailing orthodoxy" (1996: 448), and they offer an explanation of why this orthodoxy goes astray. The explanation is that failures to exercise free will are "manifest" to us for example when we are weak-willed, whereas "failures to exercise free thought . . . are essentially elusive" (1996: 449). However, it seems to me that we are aware of believing things against our better reasons just as we are aware of desiring things against our better reasons. For instance, if we are in a spooky place like a graveyard, we may experience a belief in there being ghosts around creeping in on us, though we are utterly convinced that there is conclusive evidence against such entities.

If we are not responsible for our beliefs even when they are the upshot of deliberation, why is that? A first suggestion might be that this is because the reasons upon which beliefs are based are reasons of a wrong kind: not reasons for actions, but reasons for beliefs which support the *truth* of those beliefs. However, I do not think this is quite right, since it seems that we are often not responsible for our desires, though they lack truth-value. For instance, our desires to feel pleasure, and to avoid pain, for their own sakes appear to be desires for which we cannot be responsible. It might be that this is because when we desire something (solely) for its own sake, we do not desire it for a reason.[7] But your desire to avoid a stimulus, for example, a naked flame, for the reason that, as you believe, it will

cause you pain, seems to be another desire for which you are not responsible. Similarly, your emotion of fearing the flame for the same reason seems also to be a judgment-sensitive attitude for which you are not responsible.

My belief that the traffic light will change to green, and your aversion to and fear of the painful stimulus, have in common that they are forced upon us. We cannot avoid having them, except indirectly, say, by knocking ourselves unconscious; we have no direct control over them as we have over our intentional actions. However, Scanlon denies that the fact that attitudes "arise in us unbidden" (1998: 22) is the reason why we are not responsible for them. In opposition to him, I believe that there is some truth in this view. My proposal is that what we are responsible for is *what is an outcome of practical deliberation* – that is, what is both explained and, in our own eyes, justified by consideration of reasons for action. Intentions, decisions, tryings, and intentional actions are all outcomes of practical deliberation, however fleeting. They cannot come to us "unbidden", that is, without being "invited" by consideration of action reasons. Desires and emotions may visit us uninvited by practical deliberation.[8] They may violate at least one of two necessary conditions which, I propose, together are sufficient for us to be responsible for something: (1) that it is based upon practical reasons, and (2) that it is not forced upon us, but is sensitive to consideration of further reasons. So, if you have a desire to avoid a stimulus simply because it is painful, you are not responsible for this desire, though it is based upon a practical reason. But if, in the light of further reasons, you desire this all things considered and decide and intend to avoid the stimulus, you are responsible for your desire, decision, and intention, since (2) is also satisfied.

However, a qualification should immediately be added to this account of the object of responsibility: when I deny that we are responsible for our beliefs, desires, and emotions, I mean to deny that we are *basically* responsible for them. We can intentionally acquire beliefs, desires, and emotions by performing certain actions; for instance, I can acquire beliefs about the visual appearance of something by intentionally looking at it with the goal of acquiring such beliefs. In such circumstances, we can be nonbasically responsible for acquiring these beliefs, desires, and emotions, by virtue of being responsible for the actions by means of which we acquire them. Similarly, if I am responsible for pulling the trigger of a loaded gun, I can be nonbasically responsible for the intended or foreseen effects of this event, such as the firing of this gun, the death of the victim, and so on. It is an important feature of the relation *being responsible for* that we can be said to be nonbasically responsible for the intended or foreseen effects of what we are basically responsible for. In this extended sense, we can be held responsible for all sorts of things which are anticipated consequences of that for which we are basically responsible.

It might be wondered where to locate precisely the divide between what we are basically and nonbasically responsible for. I am inclined to think that what we can be basically responsible for is all-things-considered desires, decisions,

intentions, tryings, and the simplest intentional actions, such as moving a finger, which we do not perform by means of performing other intentional actions. By contrast, if we bring about something by applying what we conceive as means to it, such as firing a gun by means of intentionally pulling a trigger, it is something for which we can be only nonbasically responsible. Our responsibility is nonbasic here because we perform these acts by means of something else for which we are responsible. This is not true of the simplest intentional actions.

When I claim that we are basically responsible for all-things-considered desires, decisions, intentions, tryings, and the simplest intentional actions, I am assuming that these items are not causally related. Rather, to decide to do something *is* to form an all-things-considered desire or intention to do it, and an all-things-considered desire or intention to do something turns into a trying to do it when the appropriate time is believed to have arrived. Finally, trying to do something, in the presence of the requisite ability to do it, *is* intentionally doing it, and if the ability does not consist in doing something by means of doing something else, the resulting act is an object of basic responsibility. But for present purposes the precise location of the line between basic and nonbasic responsibility is not crucial. What *is* important is that the concept of responsibility is such that responsibility can be extended to expected effects of what we are (basically) responsible for; otherwise, the concept of responsibility would be too restrictive to be of much use.

To say that we are morally responsible for something, X, is to say that we can appropriately or justifiably be rewarded for X, if the value of X is (highly) positive, and punished for X if the value of X is (gravely) negative. So, an account of responsibility must explain how rewarding and punishing could be justifiable. This is a requirement that my account satisfies, since the prospect of being rewarded or punished for something that we propose doing constitutes a reason for or against performing that action which could be considered in deliberation. Thereby, a reward could be justified as something that encourages us to perform beneficial actions, and a punishment could be justified as something that deters us from performing harmful actions. This is a forward-looking or consequentialist justification. It stands in contrast to the backward-looking or retributivist justification of rewards and punishment in terms of their being deserved, which will be examined in the section "Responsibility and Desert."

The forward-looking justification of the practice of rewarding and punishing is perfectly compatible with the rule of determinism in the realm of mind and action. Deliberation about whether or not to perform an action presupposes that the action is such that it is for deliberators both *epistemically* possible that they decide to perform that action (and perform it) and *epistemically* possible that they decide not to perform it (and do not perform it). That is to say, deliberators must not possess information which entails either that they decide to perform this action, or that they decide not to perform it. But this is compatible with there *in fact* being causes which necessitate their making one decision or the other, which is what determinism implies.

Although some philosophers would disagree (see, for example, van Inwagen 1983: 162–82), I think that Harry Frankfurt has convincingly shown that we could be responsible for something, even though we could not in fact have avoided it. Suppose that I decide to pull the trigger of a gun at a time, t, but that the circumstances are such that, if I had not made this decision, I would still have pulled the trigger at t because of an uncontrollable spasm. Since I do make the decision, my pulling the trigger of this gun at t is overdetermined. Nevertheless, it strikes me as clear that I am responsible for pulling the trigger. Thus, for me to be responsible for an action, it is sufficient that I perform it *because of* a decision I make; I need not perform it *only* because of the decision (cf. Frankfurt 1988: 10).

Even so, if I *knew* that, independently of my decision, there was a sufficient cause of my pulling the trigger, this would undercut my power to decide to do it. So, our capacity for deciding and, thus, acting intentionally presupposes that as a rule there are not such sufficient causes, since if there were such causes *as a rule*, we would be justified in assuming that they are probably present now, and this assumption would undermine current deliberation. However, the fact that determinism reigns in the realm of mind and action does not imply that there are any such *decision-independent* sufficient causes of our intentional actions. It implies that there are sufficient causes, but these could include our decisions (and their neural correlates) as nonredundant parts. If they do, it is, in principle, impossible for us to circumvent the outcomes of our practical deliberations by making, on the basis of knowledge about these causes, reliable predictions about what we shall do. For such predictions may themselves influence deliberation, and they are debarred from taking into account their own influence. For instance, if you predict that you will give in to a temptation, awareness of this prediction might make you defiant, so that you will no longer give in to it. Thus, a prediction might falsify itself. Of course, you might make another prediction which takes into account the effects of the first prediction, but the second prediction might also influence what it predicts. Taking this influence into account requires yet another prediction, and so on.[9]

So, responsibility, as it manifests itself in practical deliberation, is not threatened by determinism. I call this *direct* responsibility as opposed to the *ultimate* responsibility which, as will emerge in the section "Responsibility and Desert," *is* undermined by determinism (and indeterminism). Threats to direct responsibility are more local than the truth of determinism – they come from various psychic disorders, nonnegligent misinformation, and so on – and undercut the responsibility only of some of us some of the time, not all of us all the time, as determinism would do. They undercut responsibility by disrupting practical deliberation.

We are then – directly and basically – responsible for what is the outcome of our deliberations about what to do. But we also have a power to make up our minds about what to believe on the basis of theoretical deliberation. Yet, I have suggested, this does not make us responsible for what we believe. The explanation lies, I suggest, in the difference between the attitudes that reasons for beliefs and

practical reasons, respectively, are reasons for. Practical reasons are reasons for attitudes whose function it is to make the world fit their objects, while reasons for belief are reasons for an attitude whose function it is to have a content that fits the world; they are reasons that support the truth of this content.[10] The content of practical reasons describes effects of actions that we take it to be in our power to execute. Responsibility – in the nonbasic form – extends from these actions to their consequences if these consequences become actual.

Responsibility and Causation

It is a familiar fact that the expression "being responsible for" sometimes means "causes." For instance, to say that lightning was responsible for the power cut is to say that lightning caused the power cut. The fact that "responsibility" has this purely causal meaning, alongside the meaning of moral responsibility, may be an indication that the moral use has developed out of the causal use, and that causal elements are still part of the moral use. Herbert Hart and Tony Honoré put forward a similar idea when they write that this ambiguity of the term "responsibility" is "some testimony to the primacy of causal connection as an element in responsibility" (1985: 65), and that "doing or causing harm constitutes not only the most usual but the primary type of ground for holding persons responsible" (1985: 65). I shall now discuss this view that causal facts are a part of moral responsibility or of what morally justifies praise and blame, reward and punishment, so that how praise- or blameworthy people are is partly a function of what they have caused. This is what I call the role of responsibility in causally differential contexts.

Even if causing an event is necessary for being morally responsible for it, it is surely not sufficient. As should be clear from the preceding section "Responsibility and Reasons for Action," there are mental conditions that must be satisfied: our causing of something must be *intentional*, *negligent*, or some such, in order for us to be morally responsible for it. Surprisingly, however, we sometimes feel morally responsible even for what we unintentionally and nonnegligently cause. For instance, a man could feel responsible for making a woman pregnant, though he could not reasonably have known that the condom he was using was defective. According to the law of many countries, he would have to shoulder parental duties for the resulting child, pay an allowance, and so on, which evidences that the law regards him as responsible for the pregnancy (cf. McMahan 2002: 374).

This situation is akin to what Bernard Williams has called "agent-regret": in his example, a lorry driver has this feeling when he accidentally kills a child who unexpectedly darts out in front of his truck (1981: 27ff.). But Williams' term "agent-regret" strikes me as infelicitous. What Williams' lorry driver is feeling is not regret, but (unreasonable) *guilt* because he ran down and killed a child. In opposition to this view, Michael Moore maintains that what the lorry driver is

feeling is "a recognizable species of regret, felt on occasions of faultless causation of harm to others" (2009: 32). No doubt, the lorry driver may regret that he ran down the child, but it is not all that he is feeling. First and foremost, he has a crushing feeling of being guilty and deserving of blame and disapproval for having hit the child. Moreover, it seems that the example of the man with the defective condom could be used to support this view. For we have fewer scruples against classifying his emotion as guilt, and him as somehow responsible for the pregnancy, when the law imposes upon him the duties of fatherhood.

There is another kind of case which illustrates our prereflective tendency to regard causation as sufficient for moral responsibility. Some philosophers, for example Kamm (1992: 45–50) and Thomson (1990: 369–70), have followed Nozick (1974: 34–5) in thinking it permissible to kill *innocent threats* (and innocent shields) in self-defense. Nozick's example is that of using his ray gun to cause a falling heavy man to disintegrate in midair. The heavy man would otherwise crush Nozick who is at the bottom of a well into which the heavy man has been pushed, but would himself survive because Nozick would cushion his fall. These philosophers regard this killing as permissible, although they deny that it is permissible for us to kill innocent *bystanders* in self-defense, for example, to grab a bystander and use him or her as a shield against a lethal projectile. But the difference between these cases seems purely causal: while the innocent threat will cause death, the innocent bystander will not. It appears that we instinctively feel that the threat would somehow be morally responsible and blamable for the resulting death, albeit he or she is innocent. This ambivalence is not easy to make sense of.

I believe that, on reflection, most of us will recognize that the mere fact that we cause something cannot in the least make us morally responsible for it, or deprive us of our innocence as regards it. We should instead regard the feelings that I have described as disposable vestiges of the causal origin of our concept of moral responsibility. Perhaps, in the distant past, human beings considered themselves morally responsible for more or less everything that they caused. Think for instance of the strict liability of ancient tragedies – for example, that Oedipus was held morally responsible for killing his father and having sex with his mother, though he had no way of knowing that these people were his parents. But this view is now obsolete, and if we abandon it, then, so far as I can see, we must also abandon the view that there is a moral difference between the innocent threat and the innocent bystander.

What I want to focus upon instead is the much more reasonable claim that, alongside some mental condition, having caused an event is a *necessary* condition for being morally responsible and blame- or praiseworthy for its occurrence. This idea crops up in the act–omission doctrine which implies, for instance, that it is impermissible to kill one innocent individual in order to save more innocents from being killed. Something can be morally wrong only if we can be responsible for it; so if we are not at all, or only a little, responsible for letting something happen, it cannot be (very) wrong to let it happen.

I want to suggest that we do not *cause* what we let happen, by omitting to act. Suppose, for instance, that I let my hand remain where it is, resting on the table. Then my decision not to move my hand is not a cause of my hand's not moving. This is indicated by the fact that if we imagine that I had not made this decision (say, because I was absentminded or unconscious), but the circumstances in all other respects were left intact, the result would in all likelihood have been the same with respect to my hand: it would remain where it was. True, in the absence of the decision, I could not be described as *omitting* to move my hand and *letting* it remain where it is, but my hand would still remain where it was. When we let something happen, we could prevent it from happening by behaving differently, but we do not cause its happening by not acting. Thus, if I let you go on sleeping by omitting to move my hand, I do not cause you to go on sleeping, although I could have caused you to wake up by moving my hand (which I could do). We could truly say that you went on sleeping *because* I decided not to move my hand, and did not move it, but this statement explains, I submit, not by citing a cause of your going on sleeping, but by excluding causes that would have made you wake up, for example, a decision to move my hand and a consequent movement of my hand (cf. Persson 2006: 137–8; 2013a: chs. 3.2 and 5.2).

To find out whether the mere fact that you cause an event makes a difference to your moral responsibility for its occurrence, compare the following two situations:

(C1) A mechanism will cause the trigger of a gun to be pulled if you do not switch off the electrical current activating it. The gun is pointing at a victim, Vic, who will be killed if the gun fires. You let Vic be shot dead by omitting to switch off the current.

(C2) The mechanism is connected to your arm, and the electrical current will cause you a spasm that will make you pull the trigger if you do not flex some muscles in your arm. You let yourself shoot Vic dead by omitting to flex those muscles.

Instinctively, you might feel more responsible for the shooting in (C2), like an innocent threat might feel more responsible for a resulting death than an innocent bystander does. But, on reflection, this feeling is hard to sustain: surely, it cannot matter morally if the causal process leading to death is wholly external to your body, or instead partly internal to it. Therefore, the fact that in (C2) you let *yourself* shoot Vic dead does not make you more responsible and blameworthy than you would be if you had let the mechanism pull the trigger in (C1).

This makes it hard to uphold the act–omission doctrine, for compare (C2) to a situation in which you intentionally fire the gun. It is difficult to find any moral difference between these two cases: your mental state or motivation might be similar – in both cases you might want Vic to be shot dead – and in both situations you cause his death by pulling the trigger. Suppose that the only way you

could quell the spasm, which would cause your finger to pull the trigger, was by intentionally pulling the trigger. Thus, you face the choice of either letting yourself shoot Vic, by allowing the spasm to run its course, or intentionally shooting him. Imagine that his death would be somewhat less painful in the latter case. Then it seems that you ought to shoot Vic intentionally. But if so, intentionally shooting Vic could not be morally worse than letting yourself shoot him. Since letting yourself shoot him is not morally worse than letting something else, like the mechanism in (C1), shoot him, I conclude that intentionally shooting Vic is not morally worse, and does not make you more morally responsible and blamable for his death than letting him be killed. So, having caused a harmful event is not necessary for being morally responsible and blameworthy for its occurrence; you would be as morally responsible and blameworthy were you to let it occur.[11]

There is another context, in which the fact that we cause something has been thought to affect our moral responsibility, that brings us back to the notion of nonbasic responsibility discussed in the section "Responsibility and Reasons for Action." It has been thought that we are more responsible and blameworthy if we *actually* cause some harm that we try to cause, or that we knowingly risk causing, than if the harm does not result. Michael Moore is a champion of this view: "We *are* more blameworthy when we cause some evil, than if we merely try to cause it, or unreasonably risk it. The reason we *feel* so guilty in such cases is because we *are* so guilty" (2009: 30). I agree that we often feel more guilty if we actually cause harm under these circumstances, but would like to claim that this is an irrational guilt comparable to what Williams' lorry driver and the innocent threat are prone to feel. It might be easier to see this in cases in which it is quite obvious that the upshot is due to factors beyond our control. Consider the following modification of a case originally put forward by Robert Kane (1996: 55): Unaware of each other, two nuclear facility employees each puts a piece of radioactive material in a drawer of an executive's desk with the intent to kill her. Each employee knows, however, that it is not certain that the material will emit enough radioactivity to kill the executive, but that this is more probable than not. Suppose that the piece planted by employee A emits enough radioactivity to kill, and that the executive dies as a result, but that the piece planted by B does not emit any radioactivity. In this situation I am not sure whether A would feel more responsible and guilty than B if both were informed *post factum* about what has happened. The more important point is, however, that even if A were to feel more responsible and guilty, (s)he *should* not feel it because (s)he *is* not more blameworthy; both employees are equally blameworthy.[12]

To be sure, there is a sense in which only A could be said to be responsible for the executive's death and, thus, that (s)he alone can be guilty of having brought it about. Perhaps this is what (mis)leads some people, like Moore (and Kane), into thinking that A is more blameworthy. But this is an instance of somebody's being nonbasically responsible for something in virtue of being basically responsible for something else of which the former is an anticipated consequence. A is

nonbasically responsible for more harm than B, but this does not mean that A is *more blameworthy* than B. Random events which are beyond the control of the employees surely cannot affect their blameworthiness. One's degree of blame- or praiseworthiness is fixed by what one is basically responsible for; in the case of both employees this is exposing the executive to the same estimated risk of death. There are two possible senses of being "more responsible" which must be kept apart: being responsible for more events and being more blame- or praiseworthy. The former does not imply the latter; that you are nonbasically responsible for more harm does not imply that you are more blameworthy. A is nonbasically responsible for more harm, but not more blameworthy.

It should be clear that my argument for denying that the fact we happen to cause something could affect the degree of our blameworthiness is not the one that Moore summarizes as follows: "We lack control over the results of our actions so . . . our blameworthiness cannot be increased by the happenstance of such results" (2009: 434). I am not denying that we can have control over the results of our actions. We have this control, for example, in cases in which we are justifiably *certain* that we shall succeed in bringing about some result. The case of the nuclear facility employee is, however, not of this kind. But even though the agents do not fully control the result in this case, I do not deny that they could be held responsible for the executive's death if that is a result of their actions. This is implied by my acceptance of nonbasic responsibility: that we could be described as being responsible for expected effects of what we are basically responsible for. What I deny is that "the happenstance of such results," whether or not we have full control over them, could increase our *blameworthiness* (or praiseworthiness). I want to claim, in Moore's words, that "there is no (added) moral responsibility for the results of our actions"; that "we are not (more) responsible because of them" (2009: 24), in a sense that implies that we are more blameworthy. There is merely nonbasic responsibility for these results.

By contrast, what we are nonbasically responsible for could affect how much *compensation* we owe victims: if we are successful in our attempt to cause harm, we are likely to owe the victims more compensation than if we are unsuccessful. At first blush, it might seem unfair that the amount of compensation that agents owe their victims could depend upon the good or bad luck they have with respect to the materialization of the risks they take. But on closer inspection it seems quite fair that those who consciously run significant risks with the fates of others have to accept that they thereby expose themselves to the risk of having to pay compensation in accordance with the materialization of those risks.

The (false) idea that what we actually cause affects our degree of blame- and praiseworthiness has implications for collective actions. It implies that if several agents intentionally cause some harm in concert, they are less blameworthy than they would have been had they caused this harm single-handedly: they are blamable only for their own causal contribution to this harm. Again, this is a view of which Moore is a spokesman: "It matters when you increase the numbers [of agents] (and thus decrease the size of each contribution)" (2009: 71); "more

of a cause, more to blame" (2009: 72). Imagine that A's piece of radioactive material emitted only 75% of the radioactivity required to kill the executive and that B's piece emitted the other 25%. Then the implication of Moore's view would be that to a corresponding degree A is more blameworthy than B – three times as blameworthy as B for the death of the executive.

This view has paradoxical consequences. It implies that agents can eschew blameworthiness by joining up with other agents who have the same blamable intentions. Consider, for instance, Derek Parfit's "harmless torturers" (1984: 80). Each of these causes an *imperceptible* increase of the stimulation of 1,000 victims by pushing a button which is connected to 1,000 pain-producing devices, but together the 1,000 torturers cause as much pain to the 1,000 victims as they did in the bad old days when each of them operated a single pain-producing device and caused excruciating pain to a single victim by causing 1,000 increases of his or her stimulation. Surely, the torturers cannot escape blame by maintaining that now they individually cause no harm because the effect they produce on each victim is imperceptible (and 1,000 times 0 is 0). I instead claim that when agents intentionally cause harm together with other responsible agents, they are as blameworthy as if they had caused this harm single-handedly. They are equally blameworthy if in both cases they have intentions to cause their victims excruciating pain, though in one case they intend to achieve this by means of the assistance of others. Of course, this could be so only if the individual torturers know about the behavior of their colleagues, so that they could be said to literally *collaborate*.

Even if the actions of some of the collaborating agents in fact play no causal role, their moral responsibility is nevertheless equal to that of the agents whose actions do play a causal role. To see this, imagine the following situation. A machine will inflict excruciating pain on Vic if a majority of 1,000 voters vote in favor of this. In more detail, the procedure is this: All voters must first cast their votes; then the machine counts up the votes, and as soon as it has counted 501 votes in favor or against inflicting pain, it stops counting, and inflicts pain or releases Vic according to the count. The votes are randomly mixed inside the machine before they are counted, so the order in which they are cast does not determine the order in which they are counted (or the mixing). Suppose that 600 vote in favor of inflicting pain. Then the 99 voters who cast votes in favor of pain that were not counted by the machine are surely as morally responsible and blamable for the pain inflicted as are the 501 who cast the pro-votes that were counted, though the former in contrast to the latter did not causally contribute to making the machine inflict the pain. (True, they contributed to making the machine start the counting, but so did those who voted against.) Therefore, having contributed causally to the infliction of pain on Vic is not necessary for being morally responsible and blamable for the infliction of pain; all the 600 voters who voted in favor of this are equally responsible and blamable.

To conclude this section: having caused an event is not necessary for being morally responsible and blame- and praiseworthy for it; one could be (fully) responsible for it by having let it occur, having unsuccessfully tried to cause it, or

having knowingly risked causing it. Our concept of moral responsibility should shake off the last vestiges of its causal origin. This will influence the moral theory we accept: it will be a consequentialist rather than a deontological morality in the sense that it rejects the act–omission doctrine (and the related doctrine of the double effect). Such a consequentialist morality will extend the range of our responsibility and, thus, make more extensive moral demands upon us. The permission to let harm occur gives us greater space to decide autonomously what to do than we shall have if we are morally required to minimize the occurrence of harm. A consequentialist morality restricts the range in which we are morally permitted to act as we please.

Responsibility and Desert

In the section "Responsibility and Reasons for Action," I contended that a consequentialist or forward-looking justification of the practice of rewarding and punishing was defensible. Such a justification is standardly contrasted with a retributivist or backward-looking justification of this practice. According to the backward-looking justification, rewarding or punishing is justified when it is *deserved*. When individuals deserve something, they do so in virtue of some of their features. Let us call this *the basis* of desert (cf. Feinberg 1970: 59). The basis of desert is supposed to make it *just or fair* that individuals get what they deserve – call it *the return* (deserved). The return consists in a benefit or burden or, in other words, something that adds to, or detracts from, the well-being of the recipient, whereas the desert basis is, rather, something that is good or bad for others. A paradigm example of a desert basis is a responsible action. Thus, the fact that individuals are responsible for good deeds is thought to make it just that they receive something that is good for them in return, and the fact that they are responsible for bad deeds is thought to make it just that they receive something that is bad for them. Presumably, there has to be a certain balance or equivalence between the value of the desert basis and the value of the return. This balance is a complicated matter that I shall here have to set aside. It will be of no importance if I am right in my argument that the concept of desert has no application.

In the present context, the crucial claim is that the basis in virtue of which we deserve a return must be something for which we are responsible. Otherwise, it seems that our getting a return proportionate to the basis cannot be just. For example, it would not be deserved and just to punish some neonates because they cause their mothers a lot of pain while being born, and reward other neonates who cause their mothers little pain. This is because neonates are not responsible for the amount of pain they cause their mothers while being born. So we have the following claim:

(1) Making some better or worse off than others by rewards and punishments, respectively, is just or fair only if they are rewarded or punished in proportion to something for which they are responsible.

For instance, it is just that some are made better or worse off than others by rewards and punishments only if this is proportionate to the greater goodness or badness of deeds for which they are responsible. As the example of the newborns illustrates, it is not just to make individuals better or worse off, to reward or punish them, because they benefit or harm others without being responsible for it.

In the section "Responsibility and Reasons for Action," I suggested that we are basically responsible for possible outcomes of practical deliberation: all-things-considered desires, decisions, intentions, tryings, and simple intentional acts. But, if determinism is true, there are facts which causally explain why we deliberated the way we did – call these facts "responsibility-explaining facts." These facts may not include every fact that is causally necessary for our being responsible for something, X, through being necessary for our very existence – for example, such general conditions as the occurrence of the Big Bang or the presence of oxygen – but may include only facts that are causally necessary and sufficient to determine that, given our existence, we are responsible for X. They will include facts that explain why certain reasons for action came to our attention, why we have a character which made us make the decisions that we did make in the light of these reasons, and why we have abilities that allowed us to execute these decisions. It is possible that we are responsible for some of these responsibility-explaining facts in virtue of intentional actions that we performed earlier. But then there are other responsibility-explaining facts which explain why we performed these intentional actions and went through the deliberations preceding them, and the question arises to what extent we are responsible for these facts. Evidently, this regress of responsibility cannot be infinite, since we are temporally finite beings who began to exist (as subjects of responsibility) sometime in the past. Instead,

(2) The responsibility-explaining facts in virtue of which agents are (directly) responsible for whatever they are (directly) responsible are ultimately all facts for which they are not responsible, that is, agents are not *ultimately* responsible for anything.

As we trace the causes of our being responsible agents backwards in time, we shall eventually reach causes for which we are not responsible, for example, causes such as our genetic dispositions and early environmental influences. Therefore, although we can be directly responsible for our intentions, and the intentional actions which embody them, we cannot be ultimately responsible for these things, since in the end they are determined by facts for none of which we are (directly) responsible.

It should be noted that indeterminism would not enable us to be ultimately responsible, either. Suppose that, in the case of some intentions for which we are directly responsible, there is no set of responsibility-explaining facts which fully explain why we formed these intentions: this is partially undetermined, a matter of chance. This also excludes our being ultimately responsible for the forming of those intentions as, to the extent that something is a matter of chance, we cannot be responsible for it because it is beyond our control.

Now, it follows from (1) and (2) that

(3) Ultimately, agents are not responsible for anything and, so, it cannot be just or fair to make some of them better or worse off than others by rewarding or punishing them in proportion to anything.[13]

This cannot be just as it cannot be just to make some better or worse off on the basis of things for which they are not directly responsible – for example, to make some infants better or worse off than others because of the amount of pain they caused their mothers while being born. The rationale of the idea that it is just to give agents rewards or punishments whose value to them equals the value to others of the acts for which they are responsible is that the value of their acts then flows from what they are responsible for. But this rationale is undercut if their acts in the end turn out to originate in something beyond their responsibility. There is no plausibility in the idea that justice consists in an equivalence between the value of the return to *us* and the value of a contribution of ours to the world, if what produces this contribution is only *up to a point* within our responsibility. For then the value of this contribution in the end comes from something for which we are not responsible, and then it is not just that we enjoy the value of a return which equals the value of this contribution.

The question of ultimate responsibility arises when we ask what distribution of well-being to recipients is just or fair. But when in the process of deliberating we ask what reasons we have to act, we need not ask what makes us deliberate in the manner that we do. Here we adopt a forward-looking instead of a backward-looking perspective. So, the fact that we are not ultimately responsible is of no relevance in the deliberative context, but it is relevant in a distributive context in which it is asked what distribution of well-being is just or fair.

It is important to stress that the argument against ultimate responsibility is consistent with – indeed, it presupposes – responsibility in the direct sense. We are directly responsible because of the feasibility of a forward-looking justification of the practice of rewarding and punishing in terms of the beneficial consequences of this practice – the encouraging and discouraging effects of it. If our intentions and intentional actions are sensitive to the reasons which figure in our deliberations, the implementation of rewards and punishments could change our future behavior – to the benefit of most of us – by providing us with reasons to form different intentions in the future. Although unjust in the absence of ultimate responsibility, an unequal distribution of well-being can then be

morally justified in terms of the beneficial consequences of such a distribution. This is tantamount to saying that we *are* responsible in one sense – the direct sense – and this is crucial, because if we were not responsible for anything then, as remarked, there could not be anything that we morally (or rationally) ought to do.

To summarize, philosophical investigation into the nature of responsibility does not undermine morality, but there is reason to believe that it leads to a revision of the content of commonsense morality. First, by bringing out that moral responsibility for an event is not a function of the causal contribution to its occurrence, it disposes of the act–omission doctrine (as well as the related doctrine of the double effect) and, thus, presses commonsense morality in a consequentialist direction. Such a morality will impose greater demands on us. Second, the denial that there is anything that could make it just or fair that some are better or worse off than others, in conjunction with the formal principle of justice mentioned in the section "Three Ways in Which Responsibility is Relevant to Morality" (to the effect that justice requires that all are equally well-off if there is nothing to make it just that some are better off than others), implies that justice requires equality. Thus, it should not be assumed that the resulting consequentialism will reject the applicability of the concept of justice, as does utilitarianism.

However, this is not the place to detail this revision of commonsense morality (for further discussion, see Persson 2005: pt V and 2013a). The objective of the present essay has merely been to exemplify how a proper understanding of moral responsibility could have moral repercussions.

Notes

1 Note that, whereas I take praise and blame to consist in the *expression* of certain attitudes, others, like Sher (2006: ch. 6), take them to consist in the attitudes themselves.
2 A reference to benefits and harms also enters into Judith Thomson's criterion of stringency of rights (1990: 152–8).
3 Some philosophers prefer instead to speak of the *doctrine of doing and allowing* because they think that it is possible to allow something to happen not just by omitting to act, but by acting. I disagree; see Persson (2013a: ch. 3.2; 2013b).
4 Cf. what Robert Nozick calls a "utilitarianism of rights" (1974: 28).
5 For a recent discussion of responsibility in such cases, see Sher (2009).
6 Pettit and Smith talk about freedom rather than about responsibility, but I take it that our having freely formed an attitude entails our being responsible for it.
7 This is true on an internalist view, according to which reasons are desire dependent, but not necessarily on an externalist view, according to which reasons are conceptually independent of desires.
8 This is less clear on Scanlon's account of desire, according to which desires are not "a special source of motivation, independent of our seeing things as reasons" (1998: 40).

But I am strongly inclined to think that we acquire some desires before we are able to see things as reasons.

9 For recent expositions of this type of argument, see Bok (1998: 79ff.) and Persson (2005: 379–82). In contrast to me, however, Bok takes responsibility as it figures in the context of practical deliberation to be all there is to responsibility.

10 This is the so-called different "directions of fit" of beliefs and desires, see, for example, Searle (1983).

11 For a fuller discussion of this argument, see Persson (2004).

12 This is a denial of the moral relevance of what Thomas Nagel calls "consequential luck" (1979). In the section "Responsibility and Desert," another kind of moral luck, namely his "constitutive luck", is denied.

13 This argument is presaged by Henry Sidgwick, though he remarks that it leads to "such a precipice of paradox that Common Sense is likely to abandon it" (1981/1907: 284). More recently, an analogous argument has been advanced by Galen Strawson (1999). I have set forth this argument more fully elsewhere (Persson 2005: ch. 34; 2007). See also Nagel (1979).

References

Bok, Hilary (1998) *Freedom and Responsibility*, Princeton, NJ: Princeton University Press.

Feinberg, Joel (1970) "Justice and Personal Desert," in *Doing and Deserving*, Princeton, NJ: Princeton University Press, pp. 55–94.

Frankfurt, Harry (1988) "Alternate Possibilities and Moral Responsibility," reprinted in *The Importance of What We Care About*, Cambridge: Cambridge University Press, pp. 1–10.

Hart, H.L.A. and Honoré, A. (1985) *Causation in the Law*, 2nd edn, Oxford: Clarendon Press.

Kamm, Frances, M. (1992) *Creation and Abortion*, New York: Oxford University Press.

Kamm, Frances, M. (1996) *Morality, Mortality*, vol. 2, New York: Oxford University Press.

Kane, Robert (1996) *The Significance of Free Will*, New York: Oxford University Press.

McMahan, Jeff (2002) *The Ethics of Killing*, New York: Oxford University Press.

Moore, Michael S. (2009) *Causation and Responsibility*, Oxford: Clarendon Press.

Nagel, Thomas (1979) "Moral Luck," reprinted in *Mortal Questions*, Cambridge: Cambridge University Press, pp. 24–38.

Nozick, Robert (1974) *Anarchy, State, and Utopia*, New York: Basic Books.

Parfit, Derek (1984) *Reasons and Persons*, Oxford: Clarendon Press.

Persson, Ingmar (2004) "Two Act-Omission Paradoxes," *Proceedings of the Aristotelian Society* 104: pt 2.

Persson, Ingmar (2005) *The Retreat of Reason*, Oxford: Clarendon Press.

Persson, Ingmar (2006) "'Consciousness as Existence' as a Form of Neutral Monism," in *Radical Externalism*, ed. A. Freedman, Imprint Academic, pp. 128–46.

Persson, Ingmar (2007) "A Defence of Extreme Egalitarianism," in *Egalitarianism: New Essays on the Nature and Value of Equality*, eds. N. Holtug and K. Lippert-Rasmussen, Oxford: Clarendon Press, pp. 83–98.

Persson, Ingmar (2013a) *From Morality to the End of Reason: An Essay on Rights, Reasons and Responsibility*, Oxford: Oxford University Press.

Persson, Ingmar (2013b) "Omissions," in *The International Encyclopedia of Ethics*, ed. Hugh LaFollette, Oxford: Wiley-Blackwell.

Pettit, Philip and Smith, Michael (1996) "Freedom in Belief and Desire," *Journal of Philosophy* 93: 429–49.

Scanlon, T.M. (1998) *What We Owe to Each Other*, Cambridge, MA: Harvard University Press.

Scanlon, T.M. (2008) *Moral Dimensions*, Cambridge, MA: Harvard University Press.

Searle, John, R. (1983) *Intentionality*, Cambridge: Cambridge University Press.

Sher, George (2006) *In Praise of Blame*, New York: Oxford University Press.

Sher, George (2009) *Who Knew? Responsibility without Awareness*, New York: Oxford University Press.

Sidgwick, Henry (1981/1907) *Methods of Ethics*, 7th edn, Indianapolis: Hackett.

Strawson, Galen (1999) "The Impossibility of Moral Responsibility," reprinted in *What Do We Deserve?* eds. L. Pojman and O. McLeod, New York: Oxford University Press, pp. 114–24.

Thomson, Judith (1990) *The Realm of Rights*, Cambridge, MA: Harvard University Press.

Van Inwagen, Peter (1983) *An Essay on Free Will*, Oxford: Clarendon Press.

Williams, Bernard (1981) "Moral Luck," reprinted in *Moral Luck*, Cambridge: Cambridge University Press, pp. 20–39.

Part III

Normative Ethics

Chapter 10

Act-Utilitarianism

R.G. Frey

So much has been written about act-utilitarianism in recent years, both by way of criticism and response, that it is not possible in the compass of a single essay even to begin to do justice to the many sides of the various debates that, collectively, comprise contemporary discussion of the theory. What I have done, in order to give an overview of the present state of play with regard to the theory, therefore, is focus upon what I take to be the most important development in that theory today.

For simplicity's sake, I shall take act-utilitarianism to be the view that an act is right if its consequences are at least as good as those of any alternative. As given, this view is consequentialist, welfarist, aggregative, maximizing, and impersonal, and the principle of utility that it endorses sets up what I shall call the utilitarian goal.

The view is consequentialist, in that it holds that acts are right or wrong solely in virtue of the goodness or badness of their actual consequences. This view may be called act-consequentialism, or, here, for reasons of brevity, simply consequentialism. It is matters to do with consequentialism, and the conflicts that consequentialist thinking is supposed to engender with ordinary morality, that will be the focus of this essay. The view is welfarist, in that rightness is made a function of goodness, and goodness is understood as referring to human welfare.[1] (I leave aside here questions to do with the inclusion of animals within the scope of the theory, though in my view the "higher" animals are to be included.) The view is impersonal and aggregative, in that rightness is determined by considering, impersonally, the increases and diminutions in well-being of all those affected by the act and summing those increases and diminutions across persons.[2] The view is a maximizing one: a concrete formulation of the principle of utility, framed in the light of welfarist considerations, is "Always maximize net desire-satisfaction."[3]

The Blackwell Guide to Ethical Theory, Second Edition. Edited by Hugh LaFollette and Ingmar Persson.

The act-utilitarian goal, understood in the light of the above characterization, then, is to maximize human welfare. The crucial question to which this goal gives rise is how best to go about achieving it, and it has for some time now been thought by act-utilitarians, especially since R.M. Hare's *Moral Thinking* (1981), that the best way of going about maximizing human welfare overall may be to forego trying to maximize it on each occasion. It is this insight, in some form or other, that has spurred the most important developments in act-utilitarian theory today.

I

What has driven and continues to drive much of the opposition to act-utiliarianism, including virtually all the different versions of rule-utilitarianism, has been the thought that some alternative view can better account for a number of our moral intuitions. Our moral intuitions, it is said, frown upon murdering or torturing someone, upon enslaving people or using them as means, upon acting in certain contexts and so using people in certain ways for mere marginal increases in utility, all of which act-utilitarianism is supposed to (be able to) license. It is supposed to license these things because of its constituent consequentialism: if such acts were to have better actual consequences than the actual consequences of any alternative act, then the act-utilitarian would be compelled to call such acts right. And this, allegedly, conflicts with our moral intuitions or ordinary moral convictions or what some people think of as commonsense morality. Even most of the *contemporary* examples of problems with act-utilitarianism take this form; thus, act-utilitarianism is said by some to produce clashes with our moral intuitions over the rightness of our showing partiality to our own projects and concerns and to members of our family without incurring the accusation of bias or selfishness or over regarding people as separate and so not treating a benefit to one person as compensation for loss to another.

(Interestingly, those who use this manner of argument against act-utilitarianism generally never mention cases and instances where our moral intuitions and consequentialism coincide. The relative absence of such cases is very often explained by the fact that the opposing theorist already has settled intuitions about a case and has convinced himself that there is a way of making those intuitions fall out of or be compatible with his theory of the right. This yields (at least) two possible positions, on one of which rightness has nothing whatever to do with an act's consequences and on the other of which the rightness of certain acts has nothing whatever to do with an act's consequences. A third position, that the rightness of an act is a function of its consequences plus something else, for example, the intention and/or motive with which it was performed, is possible but, strictly speaking, not anticonsequentialist.)

This is familiar territory in past debates over utilitarianism generally, though it is no more settled for all that, and it raises directly the question of whether our

moral intuitions have probative force in ethics. This is an important issue in its own right, separate from the fate of any form of utilitarianism, but far too broad and complex an issue to be gone into in any detail here. But a few words on it are necessary, because the assumption of its truth has driven the urge to modify act-utilitarianism.

For those inclined to the view that moral intuitions do have probative force in ethics, the trick, as it were, has been to make it appear that certain of our intuitions are more secure than others – so secure, in fact, that we believe them to be more "correct" or "true" than any normative ethical theory. Obviously, those who adopt this line need to identify which these crucial intuitions are, and various ways of doing this have been suggested. Today, reflective equilibrium methodologies, both thick and thin, are perhaps the preferred way, though some relatively straightforward, old-fashioned intuitionists still survive.

Even with the back and forth movement between intuition and principle that reflective equilibrium methodologies involve, however, it is clear that some intuitions survive and remain intact. Thus, in *A Theory of Justice* (1971), Rawls appears to think that if a moral/political theory gave the result that slavery was justified, that would be enough to demand from us amendment or abandonment of the theory. His intuition on this score needs no revision. Other writers privilege other of their moral intuitions either about particular acts or classes of acts. Of course, the more we find people, whether in our own or another culture, differing over these crucial intuitions, the more difficulty we encounter in selecting just which the crucial ones are. Thus, reflective equilibrium methodologists on the one hand and straightforward intuitionists on the other seek ways to discount the variation in these crucial intuitions, or, at the very least, to reduce the scope and depth of variations.

(Plainly, fashion changes where moral intuitions are concerned. Whereas promising and truth-telling fell into the favored class earlier, my impression today is that they do not, or, at least, that often the favored intuitions are found elsewhere. For example, today they seem in part to turn upon political orientation. Thus, someone who is politically conservative not uncommonly puts the wrongness of abortion into the favored class, whereas political liberals are very unlikely to agree.)

Whatever the scope and depth of variations, the assumption that certain intuitions survive critical scrutiny has been the springboard from which assaults upon act-utilitarianism have nearly always begun. Equally, it has been the spur for developments in the theory, as numerous act-utilitarians have responded by trying to find ever more sophisticated ways of building into the theory, on act-utilitarian grounds, all kinds of devices that permit them to obtain results in particular cases much more in line with what are thought to be the crucial intuitions selected by critics as those which survive critical scrutiny. Even some of the patron saints of act-utilitarianism have led adherents of the theory to think along these lines. Thus, in Book IV of *The Methods of Ethics* (1874), Sidgwick is at pains to impress upon his readers that act-utilitarianism, far from being a wholly destructive force in normative ethics, can often be used to provide support for parts of commonsense

morality. Of course, it will not support all parts and to every last detail, and Sidgwick himself seeks the reform of much of it. But neither, Sidgwick implies, is act-utilitarianism going to be allowed to sweep aside just any part of commonsense morality. His view appears to be – certainly with regard, for example, to justice – that sometimes the theory must give way.

The point here is not which amendments to the theory Sidgwick favored but that he thought amending the theory an appropriate response to at least certain clashes between the application of the theory in particular cases and certain of the views that are taken to constitute commonsense morality. Implicitly, therefore, Sidgwick rejects the hard line on this issue adopted by J.J.C. Smart (Smart 1973; Williams 1973). Sidgwick, no more than anyone else, had a way of indicating just how he knew which were the privileged bits of commonsense morality, but the effect of his position was to encourage act-utilitarians to search for ways of bringing the results of their theory, in its application to particular cases, in line with the dominant moral intuitions on those cases. One result of this search was to erode what were widely thought of by act-utilitarians as strengths of their theory, namely, its simplicity and ease of operation. But this erosion was thought necessary if the theory was to avoid rejection for failing to account for those moral intuitions that were held by critics to be "just too secure" to be mistaken.

I take the failed experiment with (the different forms of) rule-utilitarianism to have been in part driven by the search for this complication in utilitarian thinking. Rule-utilitarianism, at least in all its varieties with which I am familiar, has long been known to suffer from certain types of instabilities that seem irreducibly part of the theory, but what lured utilitarians to give it a try was in no small measure the thought, played up by rule-utilitarians, that it could better account for just those moral intuitions that were held (by critics) to be the crucial ones that survived critical scrutiny. To this extent, rule-utilitarians implicitly supported using the accommodation of these moral intuitions as the test of adequacy of a normative ethical theory, though usually without any articulated defense of that test. And this in part proved their undoing. For once one begins to use certain of one's moral intuitions in this way, the question is bound to arise of why we should try to accommodate these privileged intuitions within a utilitarian structure at all. Squeezed by David Lyons' extensional equivalence argument for the collapse of rule- into act-utilitarianism on the one side (1965), and by the privileged moral intuitions on the other, with the possibility, if not actuality, that they could be better accommodated within altogether nonutilitarian positions, the various forms of rule-utilitarianism failed to bite.

II

The search for complication to act-utilitarianism, in order to bring the results of the application of the theory in particular cases in line with favored moral

intuitions with regard to those cases, has led almost universally among act-utilitarians to an indirect, split-level account of the theory, of the sort to be found in Hare's *Moral Thinking*. I take Hare as the exemplar of indirect consequentialism/act-utilitarianism because *Moral Thinking* remains perhaps the best-known statement of the view, but the shift to an indirect consequentialism has been the major response by act-utilitarians to accommodate parts, including deontological parts, of ordinary morality. I think there is something both right and wrong about this tactic by act-utilitarians.[4]

The crucial move in an indirect consequentialism/act-utilitarianism is to distinguish the level of theory from the level of practice. Hare, for example, distinguishes two levels of moral thinking, critical and intuitive: he adopts act-utilitarianism at the critical level and then uses it in order to select those guides at the intuitive or practical level by which to conduct one's life. The guides selected will be those, according to Hare, whose general acceptance will maximize utility. Given that these guides have been selected with an eye to the situations that we are likely to find ourselves in (so that fantastic examples cease to tell against a theory), then action in accordance with them, Hare claims, is likely to give us the best chance of doing the right thing – that is, of performing that act whose overall consequences are at least as good as those of any alternative.

The main effect of this distinction of levels is, at the level of practice, to make Hare's theory only indirectly consequentialist, since it bars any extensive appeal to an act's consequences at the intuitive level. In turn, the effect of this indirect consequentialism is to remove entirely the picture of act-utilitarianism at the intuitive level being applied on a case-by-case basis, which is the source of the difficulty over particular cases so far as disagreement with ordinary morality is concerned. Direct consequentialist thinking produces clashes that indirect consequentialist thinking does not.

Moreover, since we do not consult consequentialist thinking on a case-by-case basis at the intuitive level, Hare's indirect consequentialism avoids other problems that such a basis of application is supposed to produce. For example, it deals with the problem, obvious for some time now, that, in seeking to maximize human well-being on each and every occasion, we may fail to maximize human well-being overall; we may do better overall not to try to maximize well-being on each occasion. The indirect move addresses this problem. Again, since in principle there is no reason why, in Hare's hands, act-utilitarian thinking at the critical level might not yield as guides for life at the intuitive level precisely those patterns of rules or strict moral duties or schemes of full-fledged moral rights that deontologists advocate, his theory can accommodate the kinds of claims about rules, duties, and rights that might be held to be a vital part of deontological morality.

As I say, all this has been well known for some time now, and work on act-utilitarianism by theorists today nearly always begins with the move to an indirect consequentialism. This is why the continuing criticism of act-utilitarianism of the direct kind, whether by rule-utilitarians, various types of deontologists, or editors of ethics textbooks for undergraduates, falls wide of the mark. Equally, however,

it is why the striking results of act-utilitarian thinking in particular cases, of the sort Smart used to describe, are no longer characteristic of the theory. For what one does at the level of practice for Hare is to follow the simple, general rules for living that act-utilitarian thinking at the critical level has selected as those which give us the best chance of maximizing human well-being. This is not to say that one may never engage in direct consequentialist thinking; but, as we shall see, Hare thinks that various forms of human shortcomings typically get in the way of assessing acts' consequences, especially as they affect our own situations, thereby making such assessments less likely to serve as reliable guides to get us to the utilitarian goal of maximizing well-being.

It is clear, then, what Hare has done. Direct consequentialist thinking at the intuitive level can produce clashes with ordinary morality: the act which has best consequences on this occasion may not be the act held by privileged moral intuitions on the case to be the right one. A shift to indirect consequentialist thinking does not give this result; it has the act-utilitarian acting in accordance with the general rules, duties, and rights, all of which can bar direct appeal to consequences in order to determine rightness, selected by act-utilitarian thinking as giving us the best chance overall of maximizing human well-being. So, if clashes with some (privileged subset of) ordinary moral intuitions on particular cases is the issue, Hare has a way of disposing of them. Because his theory is indirectly consequentialist, we do not do our moral thinking at the intuitive or practical level in the way, stereotypically, act-utilitarians have been portrayed: deciding what to do on each and every occasion, on a case-by-case basis, by trying to determine which of the alternative acts available to us has best consequences.[5] Rather, at the intuitive level, we are to think in terms of the general rules (or whatever) that act-utilitarian thinking at the critical level has selected for us.

Foundationally, the resultant theory remains act-utilitarian: rules and other deontological considerations figure, as it were, in the superstructure of the theory. Accordingly, it is always open to an opposing theorist to object that the rules or rights embedded in the theory are not embedded foundationally, that this is what really matters so far as the views of ordinary morality is concerned, and that this is precisely what indirect strategies of consequentialist thinking still do not give us. This is unlikely to impress Hare. After all, if by appeal to rules and rights he can show that, for example, acts for mere marginal increases of utility are no longer deemed right by the act-utilitarian, why does it matter that, foundationally, rules and rights are not embedded in the theory in the way they are in rule-theories or rights-theories?

Which rules we have at the intuitive level is a function of their acceptance-utility, and Hare includes under this the cost versus the benefits of making the rules part of our character. We need simple rules that do not grow too complicated and so are easy to formulate, teach, and learn. The rules so selected can permit few exceptions, none of which, in the examples Hare gives in *Moral Thinking*, involve acting for mere marginal increases in utility. Thus, Hare clearly endorses the view that

following the rules selected by act-utilitarian thinking as giving us the best chance of doing the right or optimific thing can lead to suboptimal outcomes, something else that is at odds with the usual stereotype of the act-utilitarian.

Finally, the rules selected for use at the intuitive level are not sacrosanct, and Hare gives an account of when it is appropriate to turn to critical level thinking and reflect upon the set of rules we presently have. We do this when there appear to be conflicts between several of the rules or when there appear to be large utility losses beginning to be systematically incurred or when a sort of case arises which none of our simple, general rules cover. And he appends cautions about and constraints upon how this sort of critical thinking is to be conducted.

An obvious problem with a two-level account of moral thinking is that of seepage between the levels. If and when seepage occurs, then the set of rules selected for use at the intuitive level will be directly exposed to consequentialist thinking, so that breaking a rule for a marginal increase in utility may come back into the picture. Something like the seepage point is made by Bernard Williams (1973) in a way that I find problematic and which enables us to see another significant feature of Hare's account of moral thinking.

Williams urges that, even though we have time to do critical thinking in a particular case, we cannot expose rules appropriate to the intuitive level to (the demands of) act-utilitarian thinking at the critical level. The reason is that there is an inherent conflict between the vantage point of the intuitive morality one accepts and the vantage point of the critical morality one accepts. One is not a consequentialist at the intuitive level; one is at the critical level. Williams' point, then, is presumably that if seepage occurs, then the kind of thinking appropriate to the critical level may come to affect the kind of thinking appropriate to the intuitive level. If this is the case, if as a result of seepage one may expose the rules selected for use at the intuitive level to further act-utilitarian thinking, then it remains distinctly possible that a rule will be broken for a marginal increase in utility. In short, if seepage occurs, rule-breaking for marginal increases may occur.

So, what is to prevent seepage or the damage that seepage may cause? Hare's answer involves character and dispositions: one inculcates the guides into oneself thoroughly. One tries to make one's character of a certain sort: one containing dispositions acting upon which gives one the best chance of doing the right or optimific thing. Even motives and whole lives are treated this way by Hare. What one will have done is to have made oneself into a person whose intuitive morality is at odds with rule-breaking for marginal increases in utility. Put differently, the act-utilitarian is not constantly yearning at the intuitive level for direct consequentialist thinking in order to mount a case for rule-breaking. For not only has the indirect strategy confined that thinking to the critical level but also, even if seepage occurs, the act-utilitarian has built deeply into his or her character the traits and dispositions that give him or her the best chance of doing the optimific thing. Obviously, the indirect strategy has many advantages for the act-utilitarian.

III

Hare's central motive for the move to an indirect theory is that he thinks we are more likely to do the right, that is, optimific thing if we forego act-utilitarian thinking on a case-by-case basis. He thinks this for a number of factors, such as our lack of time in which to calculate an act's consequences, our inclination to bias and temptation, our weakness in the face of various pressures, our inability to be detached and clearheaded, our lack of factual information about situations, our tendency to emphasize self-interest and self-importance and to exaggerate the effects of acts upon ourselves, and so on. We might think of these as human shortcomings, and, unquestionably, they plague us, if not always, then at least a good deal of the time. If we are concerned with what to do, then we have good reason to avoid the kind of thinking – namely, direct consequentialist thinking – that is likely to give these factors free, or more free, rein.

Of course, in addition to these human shortcomings, there are epistemological problems about where an act's consequences begin and end. Once again, these militate against our reliance upon direct consequentialist thinking at the intuitive level and in favor of reliance upon, say, some simple, general rules.

The indirect strategy, then, has a number of advantages for the act-utilitarian, and it seems right that any plausible account of the theory must adopt such a strategy. In Hare's hands, the strategy is put in an unusual light: he claims that he does not take our moral intuitions to have probative force in ethics, but in the end the main reason we are given for endorsing the indirect strategy is that it reduces conflicts with ordinary morality. There are two issues masked here, one concerned with giving an account of the right, the other concerned with giving an account of how we are to think morally about what to do. Hare runs these together, as the very title of his book indicates, and it would not be too much of an exaggeration to say that consequentialism is for Hare both an account of the right and an account of how we are to think morally. He seems to be trying to save consequentialism as an account of how we do our moral thinking by showing why, based in consequentialist reasoning, we are not to do our moral thinking at the intuitive level in consequentialist terms.

Yet, in principle at least, another option is available: we could separate our account of moral thinking from our account of the right. In fact, this can seem the natural upshot of the indirect strategy. Indeed, this seems the natural consequence of Hare's position, in that human shortcomings and epistemological problems generally get in the way of doing the assessment of consequences in the case of particular acts. The result is that we do not do our moral thinking at the level of deciding what to do on the basis of case-by-case consequentialist reasoning but rather on the basis of guides selected by that reasoning at the critical level as likely to give us the best chance of doing the right, that is, optimific thing. So it is not so much that Hare adopts an indirect strategy for reasons of conformity with certain privileged moral intuitions about particular cases or kinds of cases,

but rather because human shortcomings and epistemological difficulties in assessing where consequences begin and end make the indirect strategy a more likely way of achieving the utilitarian goal of maximizing well-being.

In fact, a split-level position is quite compatible with rejecting consequentialism as the model for our moral thinking about what to do. For the net effect of the split-level account of moral thinking is to retain consequentialism as an account of the right but to provide excellent reasons for not doing our moral thinking at the intuitive level in consequentialist terms. In Hare's hands at least, the most distinctive feature of the agent that Hare envisages us as becoming is precisely that we do not do our moral thinking at the intuitive level in consequentialist terms.

Since it is effectively nonconsequentialist at the intuitive or practical level, the indirect strategy is compatible with quite severe deontological constraints. For it is perfectly possible that act-utilitarian thinking at the critical level will select guides for behavior at the intuitive level that, except at the level of utility-catastrophes or conflicts of guides, bar direct appeals to consequentialism. This in turn means that such thinking at the critical level may well yield the result at the intuitive level of injecting (1) person-relative principles into a theory that is person-neutral, (2) some incommensurable values into a theory of trade-offs, (3) devices that limit or bar utilitarian sacrifice and such trade-offs into a theory that has traditionally been portrayed as permitting these very things, and so on. In short, on the split-level strategy, deontological devices can make their way into act-utilitarianism because at the practical level one is effectively not thinking as a consequentialist.

To critics, of course, putting deontological devices into act-utilitarianism in this way will not still their disquiet. For the theory used at the critical level will be seen as pulling in one direction and the constraints at the practical level as pulling in another, and unless the constraints are strong enough, unless they bar appeals to consequentialist thinking altogether at the practical level, except at the level of catastrophe,[6] the theory used at the critical level may reassert itself and license utilitarian sacrifice, and so on. I take this way of putting the point to be the analog of Bernard Williams' idea that the split-level position combines two thoughts – roughly, consequentialism and deontological constraints – that are incompatible. But they do not seem incompatible at all, at least if one continues to distinguish between levels of moral thinking. One does know that consequentialist thinking is going on at the critical level and that, as we have seen, there are (comparatively rare) circumstances in which direct consequentialist thinking may be engaged in at the practical level, but this per se does not strike a blow at deontological constraints any more than a view of rights as trumps is struck down by the thought that there can arise (comparatively rare) circumstances in which a right can be set aside. The thought that I would probably stand a better chance of maximizing human welfare if I rigorously adhered to the terms of contracts I make is not *incompatible* with consequentialist/utilitarian thinking. Of course, as noted earlier, to the extent that the two levels of moral thinking cannot be kept apart, the deontological constraints incorporated into act-utilitarianism will be exposed to direct consequentialist thinking, not in comparatively rare circumstances, but all

the time. This is why the point about seepage, and the inculcation of certain traits and dispositions into our character, is so important.

If, however, the deontological devices built into act-utilitarianism are strong enough effectively to make a person-neutral theory person-relative at the practical level, then what, it may be asked, is the cash value of being an act-utilitarian? The answer, of course, is in terms of the utilitarian goal of maximizing human welfare, but that is not to the point here. Rather, the point is that the split-level strategy is quite compatible with hybrid theories, ones that are act-utilitarian at the critical level but that (can) feature all kinds of deontological devices that bar consequentialist thinking (except at the level of catastrophe) at the practical level. Such devices may not be built into act-utilitarianism in the manner that a deontologist may desire, but clashes in ordinary morality that all too frequently feature consequentialist thinking riding roughshod over deontological restrictions will have very much diminished. Importantly, these clashes will have diminished, not because we take certain moral intuitions to be privileged and developed the theory to account for them, but because in our best judgment acting in certain ways will give us the best chance of maximizing human well-being.

The problem of seepage is important, then, because it could have the effect of undermining the incorporation of deontological restrictions into act-utilitarianism – an effect that results typically from trying to do consequentialist thinking at the practical level. What hybrid theories do, then, is permit consequentialism to flourish as an account of moral thinking at the critical level but to bar consequentialism as an account of moral thinking at the intuitive level, with the result that hybrid theories typically do not allow the agent at the level of deciding what to do to think as a consequentialist. Nor is this result undone by the claim that sometimes we simply have to act on the basis of the consequences available to us and decide what to do; for, excepting the level of catastrophe, hybrid theories will still have the agent deciding what to do in the light of the deontological restrictions featured at the practical level because these remain as part of the strategy we endorse as likely to give us the best chance of maximizing human well-being.

The label "act-utilitarianism," then, is misleading for three reasons. First, it can connote that consequentialism will be employed on a case-by-case basis for deciding what to do; second, it can suggest that consequentialism specifies the kind of thinking we should employ in deciding what to do; and third, it can suggest that acts are the central focus around which we should attempt to construct an account of moral thinking about what to do. The indirect strategy involves the rejection of the first, hybrid theories involve the rejection of the second, and something along the lines of Hare's account of character traits and dispositions in effect, as we shall see, involves the rejection of the third. For what the third point is about is turning ourselves into people whose actions flow from an account of character in which the traits and dispositions that comprise character are inculcated in us overseen by the utilitarian goal of maximizing human welfare. Coincidentally, this has the result that clashes with certain privileged moral intuitions are much diminished, but that result does not require that we treat those intuitions as having

probative force. Some might think the move to a hybrid theory is motivated by an attempt better to account for certain of our moral intuitions, but that is not the case. The theory incorporates the indirect strategy and deontological restrictions in pursuit of the utilitarian goal.

IV

Consequentialism is an account of what makes right acts right, not a methodology for deciding what to do at the practical level. It cannot be refuted, therefore, by showing at the practical level that consequentialist thinking can be at odds with certain privileged moral intuitions, since hybrid theories do not have us thinking as consequentialists at that level.

We may not be able to determine on any particular occasion which act is right, consequentially; but that in no way shows that acts are not right in virtue of their consequences. Once one realizes this, that consequentialism can be true as an account of what makes right acts right even if we cannot determine (for reasons given) which act is right, then one can see that the real opponent of consequentialism is other accounts of what makes right acts right, such as the traditional Catholic view that some acts are right or wrong independently of their consequences, in themselves, by their very nature. Accordingly, consequentialism as an account of the right is not refuted by showing that its use in conducting our moral thinking about what to do may engender some conflicts with ordinary morality. Samuel Scheffler (1982), David Brink (1989), and Peter Railton (1988) all note this point.

In thinking about how we determine the adequacy of any account of the right, however, do not moral intuitions creep back into the picture here? Everything depends upon how they creep back in. If all reliance upon moral intuitions about particular cases is to be given up, then how are we to determine adequacy? But it need not be true that "all" reliance upon moral intuitions about cases is to be given up; for one can sever the claim that they have probative force from the claim that they indicate a very rough path to which an account of rightness must loosely cohere. The thought would be that if an account of rightness gives repeated instances of clashes with the intuitions of ordinary morality, if it would repeatedly call right acts or classes of acts that ordinary morality calls wrong, then this information is at least relevant to the determination of adequacy. But it is not clear that consequentialism is adequate in this situation. It tells us what makes right acts right; it does not tell us which acts are right. The old-style intuitionist would claim to know which classes of acts these were, but I take it that no consequentialist would want to make such a claim.

The whole way of thinking here about intuitions strikes me as confused. The thought seems to be, if moral intuitions are to figure in the test of adequacy of an account of the right, that they give us insight into a kind of moral reality that

is external to us and also fixed and binding. With access to this moral reality, we can "see" what is right and wrong and then use that "seeing" as a kind of moral arbiter of adequacy. But quite independently of whose "sight" of this moral reality is the "accurate" or "correct" one, it is far from obvious that there is any such reality in the first place. Certain claims about moral realism in metaethics notwithstanding, we need some reason to believe that people's moral intuitions penetrate to an external reality of fixed, abiding moral *principles* that determine completely whether classes of acts are right or wrong. To get at an account of rightness by claiming that it fails to call right certain classes of acts is to treat one's moral intuitions as if they gave us an independent check – independent of any theory – on moral reality. But there is no reason to think they provide such an independent check, not because people differ over the relevant classes of acts (though this is often true), but because this whole way of thinking is misguided. The only way our moral intuitions could serve as a check on an independently existing moral reality is if there really was such a reality and our intuitions could penetrate to it, and nothing in the usual cases presented against consequentialism has even come close to establishing such an external reality. To use one's moral intuitions as arbiters is what I find objectionable. I do not object to using them as rough guides to what is likely to give us the best chance of doing the right (i.e., optimific) thing.

Again, there is no inherent reason to think that an account of what makes right acts right needs to be practiced, in order to be adequate. Such an account is not an action guide, and it is a mistake for us to have thought that it was. This still leaves the question, it might be thought, of how widely our account of the right and our intuitions about what it would be right to do in this case can diverge, but this is still to misconstrue what consequentialism is about, as if it were something other than an account of the right. Rule-utilitarians have in the main been guilty of this confusion, since the main spur for development of the various forms of rule-utilitarianism was better to account for certain of our intuitions about which acts are right, as if consequentialism was inadequate because it formed the basis of our moral thinking about what to do and did not conform with the relevant intuitions in certain particular cases.

V

If the adequacy of an account of the right is not really the central issue here, then what is? This is the issue of how we are to conduct our moral thinking about what to do. If we are not to employ consequentialism as an action guide, then how are we to think morally about what to do? Railton marks the distinction between consequentialism as an account of the right and as a decision procedure and seeks to orient his discussion about how we are to think morally about what to do in terms of character. Hare, I think, at least if we see his split-level strategy as coming

to involve him with a hybrid theory, is in the same line of work. Whether he actually thought of his position in this light, or whether he thought that the split-level strategy was the most viable way for him to try to remain a consequentialist in practice, is not to the point. It is quite clear that a hybrid theory, in the way discussed earlier, is precisely one that marks the distinction between rightness and procedure for deciding what to do.

In Hare, as well as Railton, how we are to think morally about what to do involves appeal to the traits and dispositions of character and trying to make ourselves into persons with characters of a certain kind. We are to turn ourselves into people whose actions flow from a character in which the traits and dispositions that comprise character are inculcated in us overseen by the utilitarian goal of maximizing human welfare. Hare urges that we build deep into our characters' dispositions and, through these, principles that seem likely to foster human well-being. With such inculcation, we should typically come to feel a great reluctance to depart from the way these dispositions and principles lead us to behave and so come to feel guilt and remorse when we do depart from them. Moreover, if these dispositions and principles are embedded sufficiently deeply in us, they come to supply motives for us. Thus, if the appropriate disposition or trait has been embedded deeply enough, we come to want to tell the truth, not because of consequentialist considerations, but because truth-telling has become motive for our action. This is a Harean rendition of the deontological thought that we ought to want to come to tell the truth for its own sake. Moreover, if our character comes to reflect these deeply embedded dispositions and principles, then it will reinforce the prevention of seepage between the two levels of moral thinking. We have so deeply embedded certain dispositions into our character that we are very reluctant to behave differently and feel guilt and remorse when we do. In fact, we have brought ourselves to be persons the dispositions of whose characters have come to motivate us to behave in certain ways. Any reversion to direct consequentialist thinking is thereby even more remote.

But do we not use results in particular cases to determine which traits and dispositions to inculcate? We try to see which most assist the maximization of human welfare. But how are we to determine which traits and dispositions these are, if not by seeing in particular cases what someone's acting out of one of those traits and dispositions produces in the world? This is not giving probative force to our moral intuitions about those particular cases but merely trying to determine what the world would look like if trait A as opposed to B were displayed or acted upon; one compares past and probable results produced by acting upon A and B and decides accordingly which has the better chance of maximizing human welfare. One is not using one's moral intuitions about cases to decide rightness, but using past and probable results with respect to effect upon human welfare in order to decide which trait would likely give us a better chance to maximize human welfare. It is true that we look at the results of particular cases in order to conduct this assessment of effect upon human welfare, but we do not look at those results as having probative force with respect to rightness. I can attempt to assess past and

probable results with regard to human welfare quite separately from any thought that those results must conform with certain views of ordinary morality.

The emphasis in all this is upon making ourselves into people of a certain kind. It often has been true – it was initially true of act-utilitarianism – that the focus of how we are to think morally about what to do has been to conform our thinking to some external standard, which ordinary morality codified and which our moral intuitions gave us access to. It seems to me that this emphasis is misplaced; that the central issue is an internal one about what sort of person we make ourselves into, with our actions flowing out of our character so described. And the problem is that the decision of what sort of person to make ourselves is not a decision for which an external set of procedures or steps can be given; it is not as if making oneself into a certain sort of person, one whose actions flow out of traits and dispositions that have been inculcated in us overseen by the utilitarian goal of maximizing human welfare, can be reduced to a decision procedure. We all think about what is required in order to make ourselves into, say, good parents; but it is not as if there are three things that, if done, makes one a good parent. Rather, good parenting seems more likely achieved through instilling in someone certain character traits, in getting them to do certain things out of dispositions to behave that way, in getting them to behave habitually to look after the child's welfare, and so on. The focus is not on external conformity to something, but on the internal issue of making oneself into a certain sort of person.

The analogy of good parenting is interesting for another reason as well. It must be remembered that we inculcate into ourselves certain dispositions overseen by the utilitarian goal of maximizing human welfare, and this means that we have to be able to have our dispositions to behave in particular ways sensitive to circumstances. Thus, for good parenting, our disposition to look after the welfare of our child has to be sensitive to circumstances: in one setting we may think we best do this by denying the child what he or she wants, in another by giving the child what he or she wants. It is not as if there is a fixed set of rules and procedures that could be given – one that fixed indefinitely how one was to behave to one's child, in order to be a good parent.

With the focus upon internal development of character traits and dispositions, not conformity to external standards, moral thinking about how to act takes on a different appearance. It is much more intimately connected with making oneself into the sort of creature who behaves out of certain dispositions than into the sort of creature who acts only out of consequentialist concerns. There is a single-mindedness about the latter; a kind of blindness, that gets in the way; it is as if the consequentialist had reified a consequentialist account of moral thinking about what to do into an external standard and saw himself as condemned to conform to that standard. Being a good parent, being a good person, does not look like this, and it is a mistake to think otherwise.

A great deal more would have to be said on the issue of character even to begin to flesh out this sketch of a view of how we are to think morally about what to do, certainly, if we are to think in terms of specific character traits and dispositions

to inculcate into ourselves and conceptions of lives that we might envisage ourselves as living. But even this much indicates why Bernard Williams' claim that act-utilitarianism combines two thoughts that are incompatible – one consequentialist, the other not – falls wide of the mark. For it does not combine them in ways that have the different thoughts working on the same level, that of doing our moral thinking about what to do; and the emphasis upon character and habitual action out of certain dispositions is designed to prevent seepage between levels of moral thinking from occurring. If asked for an ironclad guarantee that seepage will not occur, none can be given; but an account of character traits and dispositions can be given that make it extremely unlikely direct consequentialist thinking will revert to being a decision procedure for the act-utilitarian.

In the end, then, the label "act-utilitarianism" can be misleading, not only because it can connote that consequentialism will be employed on a case-by-case basis for deciding what to do and because consequentialism, therefore, will be taken to form the kind of thinking in which we should engage in deciding morally what to do, but also because it can lead one to think that individual acts are the focus around which we should attempt to construct our account of moral thinking about what to do. Plainly, this last is not the case. Nor, if one thinks of good parenting as the analogy, would anyone think differently; for there is no specified series of acts that must be done in order to be a good parent, as if we could list those acts and tick off whether or not one had done them.

Notes

1 I shall not here pursue important issues in value theory. Thus, if it is welfare or well-being around which the act-utilitarian's value theory orients, is well-being to be understood subjectively or objectively? Is it to be understood in terms of states of mind or desires/preferences? If the latter, which desires/preferences are at issue – actual, future, or fully informed ones? If the last, which fully informed desires/preferences, if satisfied, count as constitutive of our well-being? Are full-informational accounts of well-being really defensible? Should we move on to what are sometimes called "objective list" views of well-being, thereby effectively making something part of a person's well-being independently of how it strikes them from within the life being lived? And does the notion of well-being actually capture what we take to be of central importance about our lives from the inside?

2 There are issues here that I shall not discuss. I shall not go into the distinction between agent-neutral and agent-relative values and the claim, increasingly heard, that values, the values of agents, are agent-relative. Again, is the act-utilitarian to focus upon the greatest total or the greatest average welfare? If the former, then the charge of failing to recognize the separateness of persons is advanced (and, arguably, has been answered), whereas if the latter is emphasized, problems to do with Derek Parfit's "repugnant conclusion" (Parfit 1984) are thought to arise. The former option is the one usually endorsed. Another charge that arises here (and to which various answers have been

given) is that an aggregative act-utilitarianism severs one from one's integrity by severing one from one's projects and concerns, since the impersonal aggregation of the increases and diminutions in well-being of all those affected by what one proposes to do may require one not to act in aid of one's own projects and concerns. Here, a further charge is sometimes laid: adopting an impersonal point of view from which to aggregate well-being can precisely fail to show partiality to those – for example, one's spouse or child – to whom one can seem bound by special moral ties.

3 I shall not consider whether something other than maximizing, such as satisficing, might be selected instead around which to orient the principle of utility, and I shall not inquire into whether it is always rational to maximize. Moreover, I shall not canvas the rather extensive literature on whether interpersonal comparisons of well-being or utility are possible or whether, if they are not, act-utilitarians will simply be left with judgments of Paretian optimality. Nor shall I bother to inquire whether it is really the case, as act-utilitarians usually assume, that judgments of intrapersonal comparisons are relatively unproblematic.

4 What is wrong with the tactic is, expressed baldly, that it still does not embrace the view that consequentialism must be rejected as an account of how we are to think morally about what to do.

5 I shall not here go into the question of how exactly we determine what the alternatives available to us at any moment are.

6 There is nothing unique about consequentialism/act-utilitarianism here in containing a catastrophe exemption. Thus, rights-theorists with a conception of rights as trumps allow for a catastrophe exemption as well. Of course, to the act-utilitarian the exemption will apply to utility-catastrophes whereas to rights-theorists the exemption will apply to rights-catastrophes. The extent to which rights-catastrophes in effect amount to utility-catastrophes I shall not discuss here.

References

Brink, D.O. (1989) *Moral Realism and The Foundations of Ethics*, Cambridge: Cambridge University Press.

Hare, R.M. (1981) *Moral Thinking: Its Levels, Method and Point*, Oxford: Clarendon Press.

Lyons, D. (1965) *Forms and Limits of Utilitarianism*, Oxford: Clarendon Press.

Parfit, D. (1984). *Reasons and Persons*, Oxford: Clarendon Press.

Railton, P. (1988) "How Thinking about Character and Utilitarianism Might Lead to Rethinking the Character of Utilitarianism," *Midwest Studies in Philosophy* 13: 398–416.

Rawls, J. (1971) *A Theory of Justice*, Cambridge, MA: Harvard University Press.

Scheffler, S. (1982) *The Rejection of Consequentialism*, Oxford: Clarendon Press.

Smart, J.J.C. (1973) "An Outline of a System of Utilitarian Ethics," in *Utilitarianism: For and Against*, J.J.C. Smart and Bernard Williams, Cambridge: Cambridge University Press, pp. 3–74.

Williams, B. (1973) "A Critique of Utilitarianism," in *Utilitarianism: For and Against*, J.J.C. Smart and Bernard Williams, Cambridge: Cambridge University Press, pp. 77–150.

Further Reading

Frey, R.G., ed. (1984) *Utility and Rights*, Oxford: Basil Blackwell.
Griffin, J. (1986) *Well-Being*, Oxford: Clarendon Press.
Kagan, S. (1989) *The Limits of Morality*, Oxford: Clarendon Press.
Mackie, J.L. (1977) *Ethics: Inventing Right and Wrong*, Harmondsworth, UK: Penguin.
Railton, P. (1984) "Alienation, Consequentialism, and the Demands of Morality," *Philosophy and Public Affairs* 13: 134–71.
Slote, M. (1985) *Commonsense Morality and Consequentialism*, London: Routledge & Kegan Paul.
Sumner, L.W. (1987) *The Moral Foundations of Rights*, Oxford: Clarendon Press.
Williams, B., ed. (1981) "Persons, Character, and Morality," in *Moral Luck*, Cambridge: Cambridge University Press, pp. 1–19.

Chapter 11

Rule-Consequentialism

Brad Hooker

Introduction

Act-consequentialists hold that what makes acts morally wrong is that some alternative act would produce better consequences. Rule-consequentialists believe that what makes acts wrong is not that they have inferior consequences but that these acts are forbidden by the rules that would produce the best consequences. (For prominent presentations of rule-consequentialism, see Berkeley 1712; Austin 1832/1995; Urmson 1953; Rawls 1955; Brandt 1959, 1967, 1979, 1988, 1989, 1996; Harsanyi 1982; Attfield 1987; Haslett 1994; Hooker 2000; Mulgan 2006; Parfit 2011.) This essay focuses on rule-consequentialism.

What Constitutes Benefit?

Rule-consequentialism holds that rules are to be evaluated in terms of how much *good* could reasonably be expected to result from them. By "good" I mean whatever has noninstrumental value. But what has noninstrumental value?

Utilitarians, who have been the most influential kind of consequentialists, believe that the only thing with noninstrumental value is utility. All utilitarians have held that pleasure and the absence of pain are at least a large part of utility. Indeed, utilitarianism is often said to maintain that pleasure and the absence of pain are the *only* things that matter noninstrumentally. Certainly, this was the official view of the classic utilitarians Jeremy Bentham (1789/1907), J.S. Mill (1861/1998), and Henry Sidgwick (1874/1907) – though in Sidgwick's

The Blackwell Guide to Ethical Theory, Second Edition. Edited by Hugh LaFollette and Ingmar Persson.

case, equality seems to have independent weight as a tie breaker (Sidgwick 1874/1907: 417).

Another view is that utility is constituted by the fulfillment of people's desires, even if these desires are for things other than pleasure. Many people, even when fully informed and thinking carefully, persistently want things in addition to pleasure. They care, for example, about knowing important truths, about achieving valuable goals, about deep personal relationships, about living their lives in broad accordance with their own choices rather than always in accordance with someone else's (Griffin 1986: pt 1; Crisp 1997: chs. 2, 3). The pleasure these things can bring is of course important. Still, human beings can care about these things in themselves, that is, in addition to whatever pleasure they bring.

This view, however, can be challenged. Some desires seem to be about things too unconnected with you for them to play a direct role in determining your good. Would your desiring that a stranger recovers fully from her illness make her recovery good for you, even if you never see or hear from her again (Parfit 1984: 494)? Naturally, the fulfillment of such a desire would *indirectly* benefit you *if* it brought you pleasure or peace of mind. But this is not to say that the fulfillment of your desire that the stranger recovers itself constitutes a benefit to you. Rather, if you get pleasure or peace of mind from the fulfillment of this desire, this pleasure or peace of mind constitutes a benefit to you (since you doubtless also desire pleasure and peace of mind for yourself).

The view that the fulfillment of your desires itself *constitutes* a benefit to you – if this view is to be at all plausible – will have to limit the desires that count. Of your desires, the only ones whose fulfillment constitutes a benefit to you are the ones whose content refers to you (Overvold 1980, 1982; Hooker 1991). Your desire that the stranger recovers does not refer to you. So her recovery does not itself constitute a benefit to you.

There seem to be reasons for further restrictions on the desires directly relevant to personal good. Think how bizarre desires can be. When we encounter particularly bizarre ones, we might begin to wonder whether the things are good simply because they are desired. Would my desire to count all the blades of grass in the lawns on the street make this good for me (Rawls 1971: 432; cf. Parfit 1984: 500; Crisp 1997: 56ff.)? Whatever *pleasure* I get from the activity would be good for me. But it seems that the *desire-fulfillment as such* is worthless in this case. Intuitively, the fulfillment of my desires constitutes a benefit to me only if these desires are for the right things (Finnis 1980, 1983; Parfit 1984: App. I; Hurka 1987, 1993; Brink 1989: 221–36; Scanlon 1993; Griffin 1996: ch. 2; Crisp 1997: ch. 3). Indeed, some things seem to be desired because they are perceived as valuable, not valuable merely because desired or pleasant (Brink 1989: 230–1, especially fn. 9).

Views holding that something benefits a person if and only if it increases the person's pleasure or desire-fulfillment are in a sense "subjectivist" theories of personal good. For these theories make something's status as a benefit depend always on the person's subjective mental states. "Objectivist" theories claim that the

contribution to personal good made by such things as important knowledge, important achievement, friendship, and autonomy is not exhausted by the extent to which these things bring people pleasure or fulfill their desires. These things can constitute benefits even when they do not increase pleasure. Likewise, they can constitute benefits even when they are not the objects of desire. Objectivist theories will typically add that pleasure is of course an objective good. These theories will also hold that ignorance, failure, friendlessness, servitude, and pain constitute harms.

For the most part, I will be neutral in this essay about which theory of personal good is best. *Usually* what gives people pleasure or enjoyment is also what satisfies their desires and involves the objective goods that could plausibly be listed. So usually we do not need to decide among these theories of personal good.

But not always. Suppose the ruling elite believe that quantity of pleasure is all that matters. Then they might feel justified in manipulating the people and even giving them drugs that induce contentment but drain ambition and curiosity, if they thought such practices would maximize aggregate pleasure. Or suppose the ruling elite believe that the fulfillment of desire is all that matters. Again, the ruling elite might feel justified in manipulating the formation of preferences and development of desires such that these are easily satisfied. Well, admittedly, our desires should be modified *to some extent* so that there is some reasonable hope of fulfilling them. But such modification could be pushed too far either in the name of maximizing pleasure or in the name of maximizing desire-fulfillment. A life could be maximally pleasurable, have maximum desire-fulfillment, and still be empty – if it lacked desires for friendship, achievement, knowledge, and autonomy.

Distribution

The term "rule-utilitarianism" is usually used to refer to theories that evaluate acts in terms of rules selected for their utility – that is, for their effects on social well-being. The term "rule-consequentialism" is usually used to refer to a broader class of theories of which rule-utilitarian theories are a subclass. Rule-consequentialist theories evaluate acts in terms of rules selected for their good consequences. Not-purely-utilitarian versions of rule-consequentialism say the consequences that matter are not limited to net effects on overall well-being. Most prominently, some versions of rule-consequentialism say that what matters is not only how much well-being results but also how it is distributed, in particular the fairness of alternative distributions.[1] Figure 11.1 might prove helpful:

Which version of rule-consequentialism is best? The problem with rule-utilitarianism is that it has the potential to be unfairly inegalitarian. Consider a set of rules which leaves each member of a smaller group very badly off and each member of a much larger group very well-off (Table 11.1). Now if no alternative rule would provide greater net aggregate benefit, then utilitarians would endorse this code.

Figure 11.1

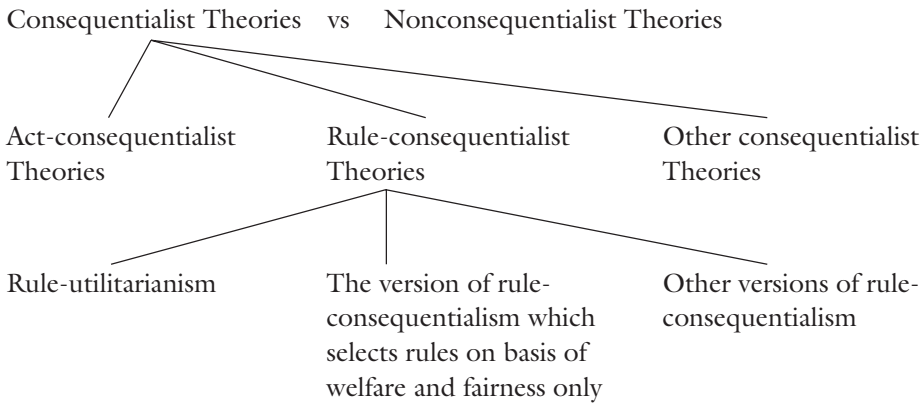


Table 11.1

Well-being

First Code: per person per group for both groups	per person	per group	for both groups
10,000 people in group A	1	10,000	
100,000 people in group B	10	1,000,000	
			1,010,000

Table 11.2

Well-being

First Code: per person per group for both groups	per person	per group	for both groups
10,000 people in group A	8	80,000	
100,000 people in group B	9	900,000	
			980,000

Yet suppose the next best rule *from the point of view of utility* would be one with the results set out in Table 11.2. Let us assume that the first code leaves the people in group A less well-off for some reason other than that these people opted to work less hard or imprudently took bad risks. In that case, the second code seems morally superior to (because fairer than) the first code. This is why we should

reject rule-utilitarianism in favor of a distribution-sensitive rule-consequentialism that considers fairness as well as well-being.

What are the relative weights given to well-being and fairness by this distribution-sensitive rule-consequentialism? Clearly, well-being does not have overriding weight. For there can be cases in which the amount of aggregate net benefit produced would not justify rules that were unfair to some group. That was what my schematic example above was meant to show.

Does fairness have overriding weight? This is particularly unsettled territory, since what exactly constitutes fairness is unclear. Nevertheless, we cannot rule out the possibility that some unfair practice so greatly increases overall well-being that the practice is justified. But it is certainly unclear where the threshold is for fairness to trump well-being. Perhaps the best we can say is that, in the choice between codes, judgment will be needed in balancing fairness against well-being. By evaluating rules in terms of two values (well-being and fairness) instead of one (well-being), distribution-sensitive rule-consequentialism is messier than rule-utilitarianism. Still, this seems to be a case where the more plausible theory is the messier one.

Criteria of Rightness versus Decision Procedures

All consequentialists, even act-consequentialists, accept that better consequences will result if we do *not* try always to decide what to do by calculating expected aggregate good than if we try always to decide in this way. Why? First, we frequently lack information about the probable consequences of various acts we might do. Where we cannot even estimate the consequences, we can hardly choose on the basis of maximizing the impartial good. Second, we often do not have the time to collect this information. Third, human limitations and biases are such that we are not accurate calculators of expected good, impartially considered. For example, most of us are biased in such a way that we tend to underestimate the harm to others of acts that would benefit us.

Now if there would be greater overall good if everyone were disposed to focus and act on *non*consequentialist considerations, then consequentialism itself would endorse these dispositions to focus and act on *non*consequentialist considerations.[2] And consequentialists nearly all agree that, in the long run and on the whole, the best consequences are likely to result from people's having firm dispositions not to harm others, not to steal, not to break promises, and so on. So rule-consequentialism and act-consequentialism do not disagree about the dispositions people should have or about the role that rules should have in guiding everyday moral decision-making.

What rule-consequentialism and act-consequentialism disagree about is the relation of such rules to the "criterion of moral rightness" (Bales 1971). Act-consequentialism holds that which behavioral dispositions people should have

and which rules people should follow in their decision making have no bearing on which acts are right or wrong – no bearing, that is, on the criterion of moral rightness.

> *Act-consequentialism* claims that an act is morally right (both permissible and required) if and only if the actual (or expected) good produced by *that particular act* would be at least as great as that of any other act open to the agent.

In contrast:

> *Rule-consequentialism* claims that an act is permissible if and only if it is allowed by a code that could reasonably be expected to result in as much good as could reasonably be expected to result from any other identifiable code.[3]

Formulations of Rule-Consequentialism

All recognizable forms of rule-consequentialism make moral rightness depend on rules evaluated in terms of their consequences. But different forms of rule-consequentialism disagree about the conditions under which rules are to be evaluated. For instance, one version of rule-consequentialism is formulated in terms of the rules the *compliance with which* would be optimific. Another version is formulated in terms of rules the *acceptance of which* would produce the most good. Should rule-consequentialism be formulated in terms of compliance or in terms of acceptance?

Although compliance with the right rules is the first priority, it is not the only thing of importance. We also care about people's having *moral concerns*. So we had better consider the costs of securing not only compliance but also adequate moral motivation. From a rule-consequentialist point of view, "moral motivation" means acceptance of moral rules. By "acceptance of moral rules," I mean a disposition to comply with them, dispositions to feel guilt when one breaks them and to resent others breaking them, and a belief that the rules and these dispositions are justified (Brandt 1967 § 8 [1992: 120–1]; 1979: 164–76, 1996: 67, 69, 145, 156, 201, 266–8, 289).

The focus on *acceptance of rules*, that is, *dispositions*, is crucial because the acceptance of a rule – or perhaps at this point it would be better to say the *internalization* of a rule – can have consequences over and above compliance with the rule (Lyons 1965: 137ff.; Williams 1973: 119–20, 122, 129–30; Adams 1976, especially p. 470; Blackburn 1985: 21 n. 12; Kagan 1998: 227–34; 2000).

The most obvious example of this involves rules that deter perfectly. Suppose you accept a rule prescribing that you retaliate against attackers. Suppose also that

you are totally transparent, in the sense that people can see exactly what your dispositions are. So everyone knows about your disposition to retaliate, and therefore *never* attacks you. Thus, your accepting that rule is so successful at deterring attack that you *never* have an opportunity to comply with the rule. Your accepting the rule thus obviously has important consequences that simply *cannot* come from your acting on the rule, since you in fact never do (Kavka 1978).

Now suppose everyone internalized rules such as "Do not kill except when killing will maximize the aggregate good," "Do not steal except when stealing will maximize the aggregate good," "Do not break your promises except when breaking them will maximize the aggregate good," and so on. Presumably, if everyone had internalized these rules, sooner or later awareness of this would become widespread. And people's becoming aware of this would undermine their ability to rely confidently on others to behave in agreed-upon ways. Trust would break down. The consequences would be terrible. And these terrible consequences would result, not from individual acts of complying with these rules, but from public awareness that the rules' exception clauses – the ones prescribing killing, stealing, and so on when such acts would maximize the good – were too available (Brandt 1979: 271–7; Harsanyi 1982: 56–61; 1993: 116–18; Johnson 1991, especially chs. 3, 4, 9).

There is another way in which a cost–benefit analysis of *internalization* is richer than a cost–benefit analysis of compliance. Getting one code of rules internalized might involve greater costs than getting another code internalized. These costs are immensely important. For example, one possible objection to a code might be that it is so complicated, or calls for so much self-sacrifice, that too much of humanity's resources would have to be devoted to getting it widely internalized. The internalization costs would be so high that internalizing this code would not, on balance, be optimal. When this is the case, rule-consequentialists hold that the code is not justified, and complying with it is not required.

Let me illustrate with an example. For each possible code we are assessing, we can juxtapose the net expected value *once the code has already been internalized* with the costs involved in the process of getting the code internalized. Suppose we are evaluating just two alternative codes, which differ in that one is simpler and less demanding than the other. Suppose the expected value of the consequences after the simpler and less demanding code has been internalized is 100 units. Suppose the expected value of the consequences after the more complicated and demanding code has been internalized is 120 units. But also suppose that the costs involved in getting the more complicated and demanding code internalized are 30 units more than the costs involved in getting the simpler and less demanding code internalized. Here we see how a cost–benefit analysis can favor the simpler and less demanding code over the more complicated and demanding one, precisely because of the difference that counting internalization costs makes.

The importance of counting internalization costs will return at a number of places in my discussion of rule-consequentialism. One such place I explore in the next section "Collapse."

Collapse

If we formulate rule-consequentialism in terms of *compliance*, we risk having rule-consequentialism collapse into act-consequentialism. The objection that rule-consequentialism collapses into extensional equivalence with act-consequentialism assumes rules are to be evaluated in terms of only the effects of compliance. While compliance can be one effect of internalizing rules, we have seen that there are also other effects. We must consider not only the benefits of compliance but also the other effects of rule internalization. With these effects factored into the evaluation of rules, the cost–benefit analysis will not favor rules extensionally equivalent to act-consequentialism.

One version of the objection that rule-consequentialism collapses into act-consequentialism claims that rule-consequentialism must favor just the one simple rule that one must always do what will maximize the good (Smart 1973: 11–12). The objection assumes that if each person successfully complied with a rule requiring the maximization of the good, then the good would be maximized. That the good would be maximized under these conditions has been challenged (see Hodgson 1967: ch. 2; Regan 1980: ch. 5).

But whether or not everyone's *complying* with the act-consequentialist principle would maximize the good, we should again consider the wider costs and benefits of rule *internalization*. The impartial good would not in fact be maximized by the internalization of just this one act-consequentialist rule. To internalize just the one act-consequentialist rule is to have just one moral disposition: the disposition to try to comply with act-consequentialism. To have just this one moral disposition is to have act-consequentialism as one's moral decision procedure. But we have already seen why act-consequentialism is not a good decision procedure.

In addition, the costs of getting internalized a disposition to try to comply with act-consequentialism would be extremely high. For getting that one rule internalized amounts to getting people to be disposed always to do what would be impartially best. Such a disposition would have to overcome people's immensely powerful natural biases towards themselves and their loved ones. To be sure, there are great benefits to be gained from getting people to care about others, and to be willing to make sacrifices for strangers. But think how much time, energy, attention, and psychological conflict would be required to get people to internalize an overriding, completely impartial altruism (if this is even possible at all). The costs of trying to make humans into saints would be too great.

That may seem like a paradoxical thing to assert. Would not a world full of people each with an overriding disposition to maximize the impartial good be so ideal as to be worth any costs of getting from here to there? I think not. Bear in mind that the costs would hardly be a once-and-for-all-time sacrifice. Rather, getting this overriding impartiality internalized would have to be done for every new generation. We are contemplating here a radical reshaping of human nature. It is not as if the impartiality internalized by one generation will be reflected in

the genes of their children. Rather, there will be the high cost of getting the overriding impartiality internalized in their children, just as there was when the parents were children themselves. (I am ignoring here the possibility of genetic engineering to create more altruistic humans.) The internalization costs will be incurred for each new generation of humans.

I have been arguing as if getting overriding impartiality internalized by the vast majority is a serious possibility, though one with prohibitive transition costs. But getting overriding impartiality internalized by the vast majority might not be a serious possibility. In any case, the only *realistic* way to make humans totally and always impartial would be to reduce their special concern for themselves and those with whom they have special attachments. What would be left might be merely a life of insipid impartiality, devoid of deep personal attachments and inimical to great enthusiasm and joy. Strong concern and commitment focused on particular projects and individuals play an ineliminable role in a rewarding human life. But these features would have to be eliminated if human beings are to internalize an overriding motivation to maximize the impartial good. (This paragraph borrows heavily from Sidgwick 1874/1907: 434; Williams 1973: 129–31; Adams 1976; Parfit 1984: 27–8, 30, 32–5; Dworkin 1986: 215; Griffin 1986: chs. 2, 4; 1996: 77, 104; Crisp 1997: 106.)

So in the light of the transition and permanent costs of getting internalized an overriding impartiality, I hold that there must be some point short of this where the costs of going further outweigh the benefits. Remember why this matters here. Getting internalized an overriding impartiality would be part of getting internalized an overriding disposition to do what will maximize the impartial good. So if there is a compelling rule-consequentialist reason against getting internalized an overriding impartiality, there is a compelling rule-consequentialist reason against getting internalized an overriding disposition to do what will maximize the impartial good. I have just argued that there is a compelling rule-consequentialist reason against getting internalized an overriding impartiality. Such a disposition would *not* find favor with rule-consequentialism. So there is a compelling rule-consequentialist reason against getting internalized an overriding disposition to do what will maximize the impartial good. This kills the first way of developing the collapse objection.

The other way of developing the collapse objection starts by admitting that internalization of just the one act-consequentialist rule would lead to bad consequences. But this way of developing the collapse objection maintains that utility could be gained from the provision of specific exception clauses to moral rules against harming others, breaking promises, and so on. If this is right, then rule-consequentialists are forced by their own criterion for rule selection to embrace rules with these exception clauses. The same sort of reasoning will militate in favor of adding specific exceptions aimed at each situation in which following some rule would not bring about the best consequences. Once all the exception clauses are added, rule-consequentialism will have the same implications for action that act-consequentialism has. This would be a fatal collapse.

To this way of developing the collapse objection, rule-consequentialists will reply by returning to the points about trust and expectations to which I alluded earlier. How much confidence would you have in others if you knew they accepted such highly qualified rules? How much mutual trust would there be in a society of agents who accepted endless exceptions to rules against harming others, breaking promises, lying, and so on?

Furthermore, the point about internalization costs is again relevant. The more plentiful and more complicated the rules to be learned, the higher the costs of learning them would be. At some point the costs of having to learn more rules, or more complications, would outweigh the benefits. Hence, the rules whose teaching and internalization would have the best results are limited in number and complexity. These limitations will keep the code from being extensionally equivalent with act-consequentialism. So this kind of rule-consequentialism does *not* collapse into act-consequentialism.

Rule-Consequentialism and the Distribution of Acceptance

A relatively simple form of rule-consequentialism selects rules by their consequences given internalization of them by 100 percent of the population. But I think the theory should be formulated in terms of internalization by less than 100 percent of the population. Rule-consequentialism needs to be formulated this way in order to make room for rules about what to do when others have no moral conscience at all. Let us refer to such people as unmitigated amoralists.

Suppose we assume internalization of the rules by 100 percent of the population. We might still need rules for dealing with noncompliance, since *internalization* by 100 percent of the people does not guarantee 100 percent *compliance*. Some people might fully accept the best rules and yet sometimes, seduced by temptation, act wrongly. Thus there is need for rules dealing with noncompliance. These rules might specify, for example, what penalties apply for what crimes. They might also specify what to do when those around you accept that they should be helping to save others but are not.

Contrast what is needed to deter or rehabilitate someone with a moral conscience too weak to ensure good behavior in some circumstances, with what is needed to deal with unmitigated amoralists (people who have no moral conscience at all). If we imagine a world with acceptance of the best code by 100 percent of the population, we have simply imagined unmitigated amoralists out of existence. Hence, we have imagined out of existence any rule-consequentialist rationale for having rules for deterring and dealing with unmitigated amoralists.

Here is why. On the rule-consequentialist view, there is always at least some cost associated with every additional rule added to the code. Every additional rule takes at least a little time to learn and at least a little memory to store. Then the question is whether there is some benefit from internalization of the rule that

outweighs the cost. We can of course frame rules applying to nonexistent situations – for example, "Be kind to any rational nonhumans living on the moon." But, if the situation envisaged really is nonexistent, where is the benefit of including such a rule in the code to be internalized? Presumably there are no benefits from such never-to-be-applied rules. These rules, which have *some* costs and *no* benefits, fail a cost–benefit analysis.

The reasoning seems to me to generate the following important conclusion: Rule-consequentialism cannot generate or justify rules about how to deter murder, rape, robbery, fraud, and so on *by unmitigated amoralists*, unless rule-consequentialism picks its rules with reference to an imagined world where there is internalization of the envisaged rules by less than 100 percent of the population. So rule-consequentialism should evaluate rules in terms of internalization by less than 100 percent of the population.

But should we assume internalization by 99 percent, or 90 percent, or 80 percent of the population, or even less? Any precise number will of course be somewhat arbitrary, but we do have some relevant factors to consider. On one hand, we want a percentage close enough to 100 percent to hold on to the idea that moral rules are for acceptance *by the whole society of human beings.* On the other hand, we want a percentage far enough short of 100 percent to *make salient the problems about noncompliance* – such problems should not be thought of as incidental. Acknowledging that any one percentage will nevertheless be somewhat arbitrary, I propose we take internalization by 90 percent of people in each future generation as the condition under which rules have to be optimal. Let me just add that this distinction between the 90 percent who are moral and the 10 percent who are amoral is supposed to cut across all other distinctions, such as distinctions in nationality and financial status.

An objection to be considered immediately, however, is whether you would be morally justified in following a rule whose internalization by 90 percent of the people in future generations would maximize expected value, even though you know that a considerably smaller percentage of people will internalize this rule.[4] As stated so far, however, the objection has not yet identified an objection. For it is very far from obvious that you are let off being required to follow a given rule merely because others are not following it. Sometimes the noncompliance of others *does* release you from being required to comply, but certainly this is not always the case. So the objection we are considering needs further development. I can think of three ways to develop it.

The first way of developing the objection is to point out that your following a rule whose internalization by 90 percent of the people in future generations would maximize expected value might in fact produce disaster, given that in fact a considerably smaller percentage of people will internalize the rule. The reply to this way of developing the objection is that one rule, among many, whose internalization by 90 percent of the people in future generations would maximize expected value, is a rule requiring you to prevent disasters when you can without excessive aggregate cost to you. I will say more about the "prevent disasters" rule later, but

the crucial point here is that the rule-consequentialist theory under discussion would never require you to do something that would produce disaster.

A second way of developing the objection is to point out that your following a rule that is costly to you and beneficial to others cannot be rightly required when those very others are not following the same rule towards you (Lyons 1965: 141). Morality cannot require us to be "suckers" open to victimization by "cheats," by which I mean those who free ride on the kindness or self-restraint of others (Mackie 1978, 1982, 1985a; Axelrod 1984).

The answer to this second way of developing the objection is that any code of rules whose internalization has high expected value will include provisions designed to enable you to put at least some kinds of pressure on others to comply with good rules towards you. One especially beneficial form of such pressure is the provision that you make your treatment on them conditional on how they treat you. For the sake of exerting this pressure on others to follow good rules, rule-consequentialism might well say that if there are people who do not follow the rules they should in their treatment of you, then morality might well let you off having to follow those rules in your treatment of them. (What the law says about such cases *might* be different because of different issues in play with the law.)

A third way of developing the objection focuses on cases where the above two conditions are not the case. These are cases where no disasters are at stake and the others towards whom one might or might not follow optimific rules are not themselves noncompliers. Suppose that the cases we are now considering are ones where, though no disasters are at stake, you could produce somewhat better consequences by following different rules from the ones whose internalization by 90 percent of those in new generations would maximize expected value. Does rule-consequentialism yield an intuitively implausible judgment about such cases? Certainly, rule-consequentialism is correct that there is much to be said for trying to establish a good precedent. (But there is more to be said here. See Hooker 2005 and 2007.)

Arguments for Rule-Consequentialism

One argument for rule-consequentialism is that general internalization of rule-consequentialism would actually maximize the impartial good. The idea is that *from a purely consequentialist point of view* rule-consequentialism seems better than act-consequentialism and all other theories.

Many act-consequentialists reply by invoking their distinction between their criterion of rightness and the decision procedure for day-to-day moral decisions. They admit act-consequentialism is not a good procedure for agents to use when deciding what to do. But they think this does not invalidate act-consequentialism's criterion of rightness. They would add that, even if rule-consequentialism is an

optimal decision procedure, this would not entail that rule-consequentialism correctly identifies what makes right acts right and wrong acts wrong.

Not only do act-consequentialists reject the consequentialist argument for rule-consequentialism, all nonconsequentialists reject it. The consequentialist argument for rule-consequentialism assumes that moralities should be assessed from a consequentialist point of view. This assumption strikes nonconsequentialists as utterly question-begging.

Another argument for rule-consequentialism is one that begins from a contractualist premise rather than a consequentialist one. The contractualist premise is that rules are morally justified just if they are ones to which everyone has good nonmoral reason to agree. The argument then proposes that an act is morally permissible if allowed by morally justified rules. Finally, the argument contends that the only rules that everyone has good nonmoral reason to agree to are the rules whose widespread acceptance would maximize expected value, impartially considered. Versions of this argument can be found in Harsanyi (1977; 1982), Brandt (1979: pt 2), Gert (1998: 215), and Parfit (2011: chs. 15–17).

Just as the first argument for rule-consequentialism assumed that moralities should be assessed from a consequentialist point of view, the second argument for rule-consequentialism assumed that moral justification is fundamentally contractualist. So let us consider other possible arguments for rule-consequentialism.

Consider the moral code whose acceptance by society would be best, that is, would maximize net good, impartially calculated. Should we not try to follow that code? Is not the code best for general adoption by the group of which we are members the one we should try to follow? These general thoughts about morality seem intuitively attractive and broadly rule-consequentialist.

And consider the related question: "What if everyone felt free to do what you are doing?" This question may in the end prove to be an inadequate test of moral rightness. But there is no denying its initial appeal. And there is no denying that rule-consequentialism is an (at least initially) appealing interpretation of the test.

Rule-consequentialism thus taps into and develops familiar and intuitively plausible ideas about morality. Morality is to be understood as a social code, a collective enterprise, something people are to pursue together. And the elements of this code are to be evaluated in terms of both fairness and the overall effects on the well-being of individuals, impartially considered.

But rule-consequentialism's leading rivals likewise emerge from attractive general ideas about morality, albeit different ones. So the fact that a theory arises from and develops attractive general ideas about morality is hardly enough to show that it is superior to all its rivals.

Now among the questions we can go on to ask about competing moral theories are (1) whether they are coherent, and (2) whether the claims they end up making about right and wrong in various circumstances are intuitively plausible. I shall not fully discuss here the objection that rule-consequentialism *incoherently* claims that maximizing the good is the overriding goal and that following certain rules can be right even when breaking them would produce more good.[5] I admit that

if we start from an overriding commitment to maximize overall good, then our rule-consequentialism might be an incoherent account of moral rightness. But our route to rule-consequentialism need not start from – and indeed most of us simply do not have – an overriding commitment to maximize overall good. If I am right about that, then the objection that rule-consequentialism is incoherent falls apart (Hooker 2000: ch. 4; 2007).

What other route to rule-consequentialism might there be? In the next few sections, I will argue that rule-consequentialism's implications about what is right or wrong in particular circumstances match our confident moral convictions quite well. But, before starting that argument, I will address the most common objection to the idea that moral theories are to be tested by their match with intuitions. This is the objection that moral convictions are merely inherited prejudices and as such cannot provide good reason for anything.

Of course people from different cultures have different moral intuitions, as do people even from the same culture. We must always be willing to reconsider our moral intuitions. They are scarcely infallible.

Although fallible, however, they can be compelling. To illustrate, compare two moral theories each of which is a coherent development of appealing general ideas about morality. Suppose one of these theories has implications that match our convictions quite closely, and the other has implications that conflict with many of our most confidently held moral convictions. In this case, I cannot see what could reasonably keep us from thinking better of the theory with the more intuitively plausible implications. Indeed, it seems to me that we are at least as confident about what is right in *some* specific kinds of situation as we are about any of the general ideas about morality that get developed into different moral theories such as virtue ethics, Kant's Categorical Imperative, moral contractualism, and act-consequentialism. This is why almost all moral philosophers are unable to resist "testing" these theories by comparing the judgments that follow from them with our confident convictions about right and wrong in various kinds of situations. But however other moral theories are defended, I am certain that appeal to reflective equilibrium between abstract theory and moral conviction must be part of the defense of rule-consequentialism.

Rule-Consequentialism on Prohibitions

Whatever act-consequentialism says about day-to-day moral thinking, act-consequentialism's criterion of moral rightness entails that *whenever* killing an innocent person, or stealing, or breaking a promise, and so on, would maximize the good, such acts would be morally right. W.D. Ross put forward the following example shown in Table 11.3 to illustrate that keeping one's promises can be right even when this would produce *slightly* less good (Ross 1930: 34–5). Most of us would agree with Ross that keeping the promise would be morally right in this

Table 11.3

Numbers below represent units of good

First Code: per person per group for both groups	Effects on person A	Effects of person B	Total good
Keeping promise to A	1,000	0	1,000
Breaking promise to A	0	1,001	1,001

case. Act-consequentialism, of course, favors breaking the promise in this case, since that is the alternative with the most good. So, if we agree with Ross about this case, we must reject act-consequentialism.

Most of us also believe (as Ross went on to observe) that, if breaking the promise would produce *much greater* good than keeping it, breaking the promise could be right. We believe parallel things about inflicting harm on innocent people, stealing, lying, and so on. Thus most of us reject what is sometimes called "absolutism" in ethics. Absolutists hold that certain acts (e.g., physical attack on the innocent, promise-breaking, stealing, lying) are *always* wrong, even when they would prevent the most extreme *disasters.*

Absolutism and act-consequentialism are, we might say, two ends of a spectrum. Whereas absolutism never permits certain kinds of act, even when necessary to prevent extreme disaster, act-consequentialism insists such acts are right not only when a great disaster is at stake but also when a *marginal* gain in net good is in the offing. Act-consequentialists seem mistaken about these cases of marginal gain, just as absolutists seem mistaken about the disaster cases. Thus, absolutism seems to go too far in one direction, act-consequentialism in the other.

Rule-consequentialism, on the other hand, concurs with our beliefs about both when we can, and when we cannot, do normally forbidden acts for the sake of the overall good. It claims that individual acts of murder, torture, promise-breaking, and so on, can be wrong even when they result in somewhat more good than not doing them would. The rule-consequentialist reason for this is that the general internalization of a code prohibiting murder, torture, promise-breaking, and so on, would clearly result in more good than general internalization of a code with no prohibitions on such acts.

Again, another rule whose general internalization would be optimal is a rule telling us to do what is necessary to prevent disasters. This rule is relevant when the only way to prevent a disaster is to break a promise or do some other normally prohibited act. In such cases, rule-consequentialism holds that the normally prohibited act should be done. I mention this rule about preventing disaster because its existence undermines the objection that rule-consequentialism would, in a counterintuitive way, prescribe sticking to rules even when this would result in disaster.

Doing Good for Others

Morality paradigmatically requires us to be willing to make sacrifices for others. Yet act-consequentialism is widely accused of going too far here too. Utility, impartially calculated, would be maximized if I gave away most of my material goods to the appropriate charities. Giving away most of my material goods is therefore required of me by (most versions of) act-consequentialism. I should probably even change to some more lucrative employment so that I would then have more money to give to charity (Unger 1996: 151). I could make much more money as a corporate lawyer, banker, stockbroker, accountant, or gossip columnist than as an employee of a philosophy department. If people should be willing to make any sacrifices that are smaller than the benefits thereby secured for others, then I should move to the better paying job so that I will have a bigger salary to contribute to the needy. With a bigger salary, I would then have to give an even larger percentage of my earnings to aid agencies. The result would be a life of devoted money making – only then to deny myself virtually all the rewards I could buy for myself with the money. After all, from an act-consequentialist perspective, my own enjoyment is insignificant compared to the very lives of those who would be saved by my additional contributions. Such reflections give special poignancy to Shelly Kagan's remark: "Given the parameters of the actual world, there is no question that [maximally] promoting the good would require a life of hardship, self-denial, and austerity" (1989: 360).

But many of us may on reflection think that it would be *morally unreasonable* to demand this level of self-sacrifice for the sake of others.[6] However praiseworthy such self-sacrifice may be, most of us are quite confident that perpetual self-impoverishment for the sake of strangers is above and beyond what morality *requires* of us.[7]

I have been discussing the objection that act-consequentialism requires us to make *huge* sacrifices in order to maximize our contribution to famine relief. Act-consequentialism also requires self-sacrifice even when the benefit to the other person is only *slightly* larger than the cost to the agent. Consider, for example, the corner office in our building. Offices are allotted on the basis of seniority. Suppose you are the most senior person who might want this corner office. But if you do not take it, it will go to an acquaintance who spends 10 percent more time in his office than you do in yours. Suppose we therefore reasonably guess that he would benefit a bit more from moving into this office than you would. This is not a life-and-death matter. Nor will he be so depressed by not getting the corner office that his work or domestic life will be seriously compromised. Nevertheless, he would get a bit more enjoyment out of the better office than you would. But you still take it for yourself. No one would think you unreasonable or immoral for doing so. Except in special circumstances, morality does not, we think, really *require* you to sacrifice your own good for the sake of slightly larger gains to others.

I have offered two objections about the demands of act-consequentialism: (1) act-consequentialism requires *huge* sacrifices from you, and (2) act-consequentialism requires you to sacrifice your own good even when the aggregate good will be only *slightly* increased by your sacrifice. In both ways, act-consequentialism is *unreasonably demanding*.

In contrast, rule-consequentialism would *not* require you to pass up the corner office and let your colleague have it. You are certainly permitted to do that if you want, but rule-consequentialism would not *require* such impartiality in your decisions about what to do with your own time, energy, money, or place in line. The rules the internalization of which could reasonably be thought to produce the most good would *allow* each person considerable partiality towards self (and even *require* partiality towards friends and family – see Brandt 1989: 100, n. 22). For, as I noted earlier, the costs of getting a complete impartiality internalized by each new generation would be prohibitive.

Likewise, whereas act-consequentialism requires huge sacrifices for the sake of maximizing the good, rule-consequentialism seems not to require more than a reasonable amount of sacrifice for this purpose. Why? A rule-consequentialist might point out that if everyone relatively well-off in the world were to contribute quite modest amounts to the best aid agencies, the worst elements of poverty could be overcome.

A rule-consequentialist will be interested in redistribution beyond what is required to secure the very basic necessities. But even after including these other potential benefits in the cost–benefit analysis, we might well conclude that the amount the world's relatively well-off would each be required to give would not be unreasonably severe. (For challenges to this view, see Carson 1991; Mulgan 1994, 1997.)

Consider the following example. Walking along a deserted road on your way to the airport for a flight to the other side of the world, you see a child drowning in a shallow pool beside the road. You could easily save the child, at no risk to yourself. But if you do save the child, you will miss your flight and lose the cost of the nonrefundable ticket.[8]

Everyone agrees you are obligated to save the child. This is true even if you are not terribly rich. Suppose the ticket costs as much as a tenth of your annual income. You would still be morally wrong not to make the sacrifice and save the child. And even if the probability of the child's drowning without your rescue is less than 100 percent – suppose, for example, it is 80 percent – you are obligated to sacrifice your ticket to save the child.

Now consider a variant of the example (Singer 1972: 233; Kagan 1991: 924–5; and Murphy 1993: 291). You and I are walking to the airport when we see two small children drowning in a lake. You and I could each easily save the children, at no risk to ourselves. The two children are positioned in the lake in such a way that you and I could each save one and still get to our flight. But if one of us saves both children, the other will miss the flight. Suppose you save one child, but I do nothing. Surely, you should now save the other.

Yet, were rule-consequentialism framed in terms of 100 percent compliance, how could it tell you to save the other? With 100 percent compliance, there would be no need for you to save the second child. With 100 percent compliance, once you had done your share, you would have done all that was needed. The rule that would be best, given 100 percent compliance, would presumably not require you to sacrifice more than you would have to sacrifice if everyone did their part. But if this rule is applied to our case, where I am in fact not coming to the rescue, you are *not* obligated to save this child. This is clearly an implausible implication.

But I argued that rule-consequentialism should be framed in terms of less than 100 percent compliance. If rule-consequentialism is framed in terms of 90 percent compliance, we can envisage that there is a need for rules about how to act when others around you are not doing their part. The rule might be, "When you happen to be surrounded by others who are not helping, then prevent disaster even if this involves doing more than you would have to do if the others helped." This rule *would* require you to save the second child from the shallow pond.

But if the world we live in – the real world – is one where partial compliance is ubiquitous, then a rule requiring you to make up for the noncompliance of others could become unreasonably demanding. Just how much would rule-consequentialism require you to make up for noncompliance by other people in a position to help? Bear in mind that Oxfam's petitioning the rich to help the very poor is hardly the only situation where some people have an opportunity to help others at relatively little cost to themselves. There will be situations where the rich can help other rich, situations where poor can help other poor, even situations where the poor can help the rich. And there will be situations where the help needs to be in the form of physical effort; other situations where the help needs to be in the form of money or time.

Given all this, perhaps the optimific rule for such a world would be: "People should help others in great need when they can do so at modest cost to themselves, cost being assessed aggregatively, not iteratively" (Cullity 1995: 293–5). Such a rule would apply in a wide array of situations – indeed, whenever some person can help another in great need. It is limited neither merely to what the rich should do to help others nor merely to what should be done concerning world poverty.

But because cost to the agent is to be assessed aggregatively rather than iteratively, the rule does not require one to help another in great need whenever the cost of helping *on that particular occasion* is modest. Having to help others whenever doing so *on that occasion* involves modest cost could easily be very costly. For each of us faces an indefinitely long string of such occasions, because any day on which we could give money to UNICEF or Oxfam counts as such an occasion. But many small sacrifices added together can amount to a huge sacrifice. The end of that road is self-impoverishment. If I am right, rule-consequentialism instead endorses a rule requiring sacrifices over the course of your life that add up to something significant. It allows but does not require personal sacrifice beyond this point.

I propose that this rule *would* have good consequences even in possible worlds that are either much poorer or much richer than ours. I do not have space here to argue either that rule-consequentialism would indeed end up with this rule in *all* possible worlds, or that this rule *always* has intuitively acceptable consequences. (For more extended discussion, see Hooker 2000: ch. 8; 2007.)

Conclusion

I have tried to fine-tune rule-consequentialism's formulation. I have also argued here that the theory develops from appealing general beliefs about morality, that it does not collapse into act-consequentialism, and that it coheres well with our intuitions about moral prohibitions and permissible partiality. Yet, even if the theory is healthy now, it is hardly invulnerable. Like someone walking through a dangerous city who has so far managed to fight off muggers emerging from behind every corner, the theory might meet an ambush it cannot survive. I am curious to see whether that happens.[9]

Notes

1 The use of "consequentialism" and "utilitarianism" so that consequentialism allows for a concern for distribution in a way that utilitarianism does not is very common. For some examples, see Mackie (1977: 129, 149); Scanlon (1978, especially s. 2); Scheffler (1982: 26–34, 70–9); Parfit (1984: 26); Griffin (1992: 126; 1996: 165). Examples of writers' including distributive considerations *within* utilitarianism include Brandt (1959: 404, 426, 429–31); Raphael (1994: 47); Skorupski (1995: 54); and arguably Mill (1861/1998).

2 See Mill (1861/1998: ch. II); Sidgwick (1867/1907: 405–6, 413, 489–90); Moore (1903: 162–4); Bales (1971); Hare (1981); Railton (1984: 140–6, 152–3); Parfit (1984: 24–9); Brink (1989: 216, 256–61). For criticism of the mileage some have tried to get from this, see Johnson (1989); Griffin (1992: 123–4, 1996: 104–5). See also Williams (1973: 123).

3 I have taken the liberty here of formulating the theory in what seems to me the most attractive relatively succinct way. I shall have to add complications to this formulation later.

4 For this objection, see Kagan (1998: 230ff.). For a proposal of how to answer it, see Ridge (2006), and for an endorsement of Ridge's solution, see Parfit (2011: 319, 419, 469). For an attack on Ridge's solution, see Hooker and Fletcher (2008). For other discussion of the issues here, see Mulgan (2006: 148–9) and Carter (2009).

5 See Lyons (1965: ch. IV); Williams (1972: 99–102, 105–8); Slote (1992: 59); Raphael (1994: 52); Scarre (1996:125–6). Relatedly, Regan (1980: 209) complains that rule-consequentialists "are only half-hearted consequentialists."

6 See Crisp (1992); and Quinn (1993: 171): "We think there is something morally amiss when people are forced to be farmers or flute players just because the balance of social needs tips in that direction. Barring great emergencies, we think people's lives must be theirs to lead."

7 Even Kagan, whose book goes on to defend act-consequentialism for 400 pages, starts by acknowledging that it "strikes us as outrageously extreme in its demands" (1989: 2).

8 This example has been central to the contemporary philosophical debate about beneficence. The example and the debate owe their prominence to Singer (1972) (see his restatement in ch. 8 of Singer 1993). Some important contributions to this discussion are Fishkin (1982); Scheffler (1982); Kagan (1989, especially 3–4, 231–2); Nagel (1991); Murphy (1993); Cullity (1994); and Unger (1996).

9 On versions of this essay, I have had help from Roger Crisp, Jonathan Dancy, Max de Gaynesford, Hanjo Glock, Hugh LaFollette, Andrew Mason, Elinor Mason, Dale Miller, David Oderberg, Derek Parfit, Ingmar Persson, and the audience at the 1997 International Society for Utilitarian Studies Conference. The present version of the essay is somewhat revised from the version published in the first edition of *The Blackwell Guide to Ethical Theory* (2000).

References

Adams, R.M. (1976) "Motive Utilitarianism," *Journal of Philosophy* 73: 467–81.

Attfield, Robin (1987) *A Theory of Value and Obligation*, London: Croom Helm.

Austin, J. (1832/1995) *The Province of Jurisprudence Determined*, ed. W. Rundle, Cambridge: Cambridge University Press.

Axelrod, R. (1984) *The Evolution of Cooperation*, New York: Basic Books.

Bales, R.E. (1971) "Act-Utilitarianism: Account of Right-Making Characteristics or Decision-Making Procedure?" *American Philosophical Quarterly* 8: 257–65.

Bentham, Jeremy (1789/1907) *An Introduction to the Principles of Morals and Legislation*, Oxford: Clarendon Press.

Berkeley, G. (1712) *Passive Obedience, or the Christian Doctrine of Not Resisting the Supreme Power, Proved and Vindicated upon the Principles of the Law of Nature*. Reprinted in D.H. Monro, ed. (1972) *A Guide to the British Moralists*, London: Fontana, pp. 217–27.

Blackburn, Simon (1985) "Errors and the Phenomenology of Value," in *Morality and Objectivity, A Tribute to J. L. Mackie*, ed. T. Honderich, London: Routledge & Kegan Paul, pp. 1–22.

Brandt, R.B (1959) *Ethical Theory*, Englewood Cliffs, NJ: Prentice Hall.

Brandt, R.B. (1967) "Some Merits of One Form of Rule-Utilitarianism," *University of Colorado Studies in Philosophy* 3: 39–65. Reprinted in Brandt (1992: 111–36).

Brandt, R.B. (1979) *A Theory of the Good and the Right*, Oxford: Clarendon Press.

Brandt, R.B. (1988) "Fairness to Indirect Optimific Theories in Ethics," *Ethics* 98: 341–60. Reprinted in Brandt (1992: 137–57).

Brandt, R.B. (1989) "Morality and Its Critics," *American Philosophical Quarterly* 26: 89–100. Reprinted in Brandt (1992: 73–92).

Brandt, R.B. (1992) *Morality, Utilitarianism, and Rights*, New York: Cambridge University Press.
Brandt, R.B. (1996) *Facts, Values, and Morality*, New York: Cambridge University Press.
Brink, David O. (1989) *Moral Realism and the Foundations of Ethics*, New York: Cambridge University Press.
Carson, Thomas (1991) "A Note on Hooker's Rule Consequentialism," *Mind* 100: 117–21.
Carter, Alan (2009) "Is Utilitarian Morality Necessarily Too Demanding?" in *The Moral Problem of Demandingness*, ed. Tim Chappell, Palgrave Publishers, pp. 163–84.
Crisp, Roger (1992) "Utilitarianism and the Life of Virtue," *Philosophical Quarterly* 42: 139–60.
Crisp, Roger (1997) *Mill on Utilitarianism*, London: Routledge.
Cullity, Garrett (1994) "International Aid and the Scope of Kindness," *Ethics* 105: 99–127.
Cullity, Garrett (1995) "Moral Character and the Iteration Problem," *Utilitas* 7: 289–99.
Dworkin, Ronald (1986) *Law's Empire*, Cambridge, MA: Harvard University Press.
Finnis, John (1980) *Natural Law and Natural Rights*, Oxford: Clarendon Press.
Finnis, John (1983) *Fundamentals of Ethics*, New York: Oxford University Press.
Fishkin, James (1982) *The Limits of Obligation*, New Haven, CT: Yale University Press.
Gert, Bernard (1998) *Morality*, New York: Oxford University Press.
Griffin, James (1986) *Well-Being: Its Meaning, Method and Moral Importance*, Oxford: Clarendon Press.
Griffin, James (1992) "The Human Good and the Ambitions of Consequentialism," *Social Philosophy and Policy* 9: 118–32.
Griffin, James (1996) *Value Judgement: Improving Our Ethical Beliefs*, Oxford: Clarendon Press.
Hare, R.M. (1981) *Moral Thinking: Its Method, Levels, and Point*, Oxford: Clarendon Press.
Harsanyi, John (1982) "Morality and the Theory of Rational Behaviour," in *Utilitarianism and Beyond*, eds. A. Sen and Bernard Williams, Cambridge: Cambridge University Press, pp. 39–62. Reprinted from *Social Research* 44 (1977).
Harsanyi, John (1993) "Expectation Effects, Individual Utilities, and Rational Desires," in *Rationality, Rules, and Utility: New Essays on the Moral Philosophy of Richard Brandt*, ed. Brad Hooker, Boulder, CO: Westview Press, pp. 115–26.
Haslett, David (1994) *Capitalism with Morality*, New York: Oxford University Press.
Hodgson, D.H. (1967) *Consequences of Utilitarianism*, Oxford: Clarendon Press.
Hooker, Brad (1991) "Mark Overvold's Contribution to Philosophy," *Journal of Philosophical Research* 26: 333–44.
Hooker, Brad (2000) *Ideal Code, Real World: A Rule-Consequentialist Theory of Morality*, Oxford: Clarendon Press.
Hooker, Brad (2005) "Reply to Arneson and McIntyre," *Philosophical Issues* 15: 264–81.
Hooker, Brad (2007) "Rule-Consequentialism and Internal Consistency: A Reply to Card," *Utilitas* 19: 514–19.
Hooker, Brad and Fletcher, Guy (2008) "Variable versus Fixed-Rate Rule-Utilitarianism," *Philosophical Quarterly* 58: 344–52.
Hurka, Thomas (1987) "The Well-Rounded Life," *Journal of Philosophy* 84: 707–26.
Hurka, Thomas (1993) *Perfectionism*, New York: Oxford University Press.

Johnson, Conrad (1989) "Character Traits and Objectively Right Action," *Social Theory and Practice* 15: 67–88.
Johnson, Conrad (1991) *Moral Legislation*, New York: Cambridge University Press.
Kagan, Shelly (1989) *The Limits of Morality*, Oxford: Clarendon Press.
Kagan, Shelly (1991) "Replies to My Critics," *Philosophy and Phenomenological Research* 51: 924–5.
Kagan, Shelly (1998) *Normative Ethics*, Boulder, CO: Westview Press.
Kagan, Shelly (2000) "Evaluative Focal Points," in *Morality, Rules, and Consequences*, eds. Brad Hooker, Elinor Mason, and Dale Miller, Edinburgh: Edinburgh University Press, pp. 134–55.
Kavka, Gregory (1978) "Some Paradoxes of Deterrence," *Journal of Philosophy* 75: 285–302.
Lyons, David (1965) *Forms and Limits of Utilitarianism*, Oxford: Clarendon Press.
Mackie, J.L. (1977) *Ethics: Inventing Right and Wrong*, Harmondsworth, UK: Penguin.
Mackie, J.L. (1978) "The Law of the Jungle: Moral Alternatives and the Principles of Evolution," *Philosophy* 53: 455–64. Reprinted in Mackie (1985b: 120–31).
Mackie, J.L. (1982) "Co-operation, Competition, and Moral Philosophy," in *Cooperation and Competition in Humans and Animals*, ed. A. M. Colman, Wokingham, UK: Van Nostrand Reinhold, pp. 271–84. Reprinted in Mackie (1985b: 152–69).
Mackie, J.L. (1985a) "Norms and Dilemma," in *Mackie* (1985b: 234–41).
Mackie, J.L. (1985b) *Persons and Values: Selected Papers*, vol. 2, eds. Joan and Penelope Mackie, Oxford: Clarendon Press.
Mill, J.S. (1861/1998) *Utilitarianism*, ed. Roger Crisp, Oxford Unversity Press.
Moore, G.E. (1903) *Principia Ethica*, Cambridge: Cambridge University Press.
Mulgan, T. (1994) "Rule Consequentialism and Famine," *Analysis* 54: 187–92.
Mulgan, T. (1997) "One False Virtue of Rule Consequentialism and One New Vice," *Pacific Philosophical Quarterly* 77: 362–73.
Mulgan, T. (2006) *Future People*, Oxford: Oxford University Press.
Murphy, Liam (1993) "The Demands of Beneficence," *Philosophy and Public Affairs* 22: 267–92.
Nagel, Thomas (1991) *Equality and Partiality*, New York: Oxford University Press.
Overvold, Mark (1980) "Self-Interest and the Concept of Self-Sacrifice," *Canadian Journal of Philosophy* 10: 105–18.
Overvold, Mark (1982) "Self-Interest and Getting What You Want," in *The Limits of Utilitarianism*, eds. H.B. Miller and W.H. Williams, Minneapolis, MN: University of Minnesota Press, pp. 186–94.
Parfit, Derek (1984) *Reasons and Persons*, Oxford: Clarendon Press.
Parfit, Derek (2011) *On What Matters*, Oxford: Oxford University Press.
Quinn, Warren (1993) *Morality and Action*, New York: Cambridge University Press.
Railton, Peter (1984) "Alienation, Consequentialism, and the Demands of Morality," *Philosophy and Public Affairs* 13: 134–71.
Raphael, D.D. (1994) *Moral Philosophy*, 2nd edn, Oxford: Oxford University Press.
Rawls, John (1955) "Two Concepts of Rules," *Philosophical Review* 64: 3–32.
Rawls, John (1971) *A Theory of Justice*, Cambridge, MA: Harvard University Press.
Regan, D. (1980) *Utilitarianism and Co-operation*, Oxford: Clarendon Press.
Ridge, Michael (2006) "Introducing Variable-Rate Rule-Utilitarianism," *Philosophical Quarterly* 56: 242–53.

Ross, W.D. (1930) *The Right and the Good*, Oxford: Clarendon Press.
Scanlon, T.M. (1978) "Rights, Goals and Fairness," in *Public and Private Morality*, ed. S. Hampshire, Cambridge: Cambridge University Press, pp. 93–111.
Scanlon, T.M. (1993) "Value, Desire, and Quality of Life," in *The Quality of Life*, eds. M. Nussbaum and A. Sen, Oxford: Clarendon Press, pp. 185–200.
Scarre, Geoffrey (1996) *Utilitarianism*, London: Routledge.
Scheffler, S. (1982) *The Rejection of Consequentialism*, Oxford: Clarendon Press.
Sidgwick, Henry (1874/1907) *Methods of Ethics*, 7th edn, London: Macmillan.
Singer, Peter (1972) "Famine, Affluence, and Morality," *Philosophy and Public Affairs* 1: 229–43.
Singer, Peter (1993) *Practical Ethics*, 2nd edn, Cambridge: Cambridge University Press.
Skorupski, John (1995) "Agent-Neutrality, Consequentialism, Utilitarianism . . . A Terminological Note," *Utilitas* 7: 49–54.
Slote, Michael (1992) *From Morality to Virtue*, New York: Oxford University Press.
Smart, J.J.C. (1973) "Outline of a System of Utilitarian Ethics," in *Utilitarianism: For & Against*, J.J.C. Smart and Bernard Williams, Cambridge: Cambridge University Press, pp. 3–74.
Unger, Peter (1996) *Living High and Letting Die: Our Illusion of Innocence*, New York: Oxford University Press.
Urmson, J.O. (1953) "The Interpretation of the Philosophy of J. S. Mill," *Philosophical Quarterly* 3: 33–9.
Williams, Bernard (1972) *Morality: An Introduction to Ethics*, New York: Harper & Row.
Williams, Bernard (1973) "A Critique of Utilitarianism," in *Utilitarianism: For & Against*, J.J.C. Smart and Bernard Williams, Cambridge: Cambridge University Press, pp. 77–150.

Further Reading

Dancy, Jonathan (1981) "On Moral Properties," *Mind* 90: 367–85.
Dancy, Jonathan (1983) "Ethical Particularism and Morally Relevant Properties," *Mind* 92: 530–47.
Dancy, Jonathan (1993) *Moral Reasons*, Oxford: Blackwell.
Foot, Philippa (1985) "Utilitarianism and the Virtues," *Mind* 94: 196–209.
Murphy, Liam (1997) "A Relatively Plausible Principle of Benevolence: A Reply to Mulgan," *Philosophy and Public Affairs* 26: 80–6.
Scanlon, T.M. (1982) "Contractualism and Utilitarianism," in *Utilitarianism and Beyond*, eds. A. Sen and B. Williams, Cambridge: Cambridge University Press, pp. 103–28.

Chapter 12

Nonconsequentialism

F.M. Kamm

Introduction

Nonconsequentialism is a normative ethical theory which denies that the rightness or wrongness of our conduct is determined solely by the goodness or badness of the consequences of our acts or the rules to which those acts conform. It does not deny that consequences can be a factor in determining the rightness of an act. It does insist that even when the consequences of two acts or act types are the same, one might be wrong and the other right. This theory denies both act- and rule-consequentialism, understood as holding that the right act or system of rules is the one that maximizes good consequences as determined by an impartial calculation of goods and bads. This calculation requires that we have a theory of what is good; it may be extremely liberal, holding that killings are bad or that autonomy is good, but we are still required to maximize the good.

Despite the name "consequentialism," many consequentialists think that we ought to maximize the goodness of states of affairs where this includes the act itself and its consequences. Nonconsequentialists deny this. Because of the possibility of this alternative contrast, consequentialism can be thought of as a form of teleology and nonconsequentialism as a form of deontology.

Contemporary nonconsequentialism finds its spiritual roots in the work of Immanuel Kant and W.D. Ross. Nonconsequentialists are drawn to Kant's formulation of the Categorical Imperative, which specifies that we should always treat rational humanity in oneself and in others as an end in itself and never merely as a means, and to his distinction between perfect and imperfect duties. Ends-in-themselves are said to have unconditional value, value independent of serving

The Blackwell Guide to Ethical Theory, Second Edition. Edited by Hugh LaFollette and Ingmar Persson.

anyone's personal ends and independent of being in a particular context. Merely counting each person's interests in the way consequentialists do is not enough to express the fact that each person is an end-in-itself. Rather, if I am an end-in-myself, then this constrains conduct that would maximize overall good.

Some nonconsequentialists suggest that we divide the imperative into two components: (1) treat persons as ends in themselves; and (2) do not treat them as mere means. If we treat people as mere means, we do not treat them as ends-in-themselves, because we are interested in them only as causally efficacious tools to some goal not serving their own ends. Nonetheless, we might fail to treat people as ends-in-themselves, even though we do not treat them as mere means. For example, if we harm someone as a *foreseen* effect of failing to constrain our conduct (Quinn 1993).

The second element of Kant's legacy that appeals to nonconsequentialists is his distinction between perfect and imperfect duties. He is thought of as an absolutist because he held that though we have some moral "leeway" in how or when we fulfill such imperfect duties as aiding, perfect duties such as not killing the innocent must always be carried out. Contemporary nonconsequentialists often deny the absolutist conception of perfect duties, but do accept that the class of negative duties (e.g., not to harm) is more stringent than the class of positive duties (e.g., to aid).

W.D. Ross was the second inspiration for contemporary nonconsequentialism. Although Ross thought there is a prima facie duty of beneficence, he also thought there are numerous other prima facie duties; for example, a duty not to harm, a duty of gratitude, a duty to do justice. If these prima facie duties conflict, as he thought they might, we have no single scale on which to weigh them, or rule by which to order them, so as to determine our actual duty. Some contemporary nonconsequentialists have tried to strengthen Ross's view by more precisely determining the relative weights or ordering of prima facie duties, or at least by more precisely characterizing them. This might require stating duties so that they specify their own limits, or finding more basic prima facie duties than Ross described that do not as easily come into conflict with each other.

Contemporary Nonconsequentialism Outlined

Nonconsequentialism is now typically thought to include personal prerogatives not to maximize the good and constraints on producing the good. A prerogative denies that agents must always maximize good consequences. Hence, it allows for the possibility that some acts are supererogatory because, while they are not morally required, they are morally valuable in virtue of producing better consequences. Constraints limit what we may do to nonconsenting people, who pose no threat and are not responsible for threats, in pursuit of our own or even

impartial good. Partial nonconsequentialists might advocate prerogatives but no constraints (Scheffler 1982) or constraints but no prerogatives (Kagan 1989).

The most commonly proposed constraints are: (1) a strong duty not to harm owed to each individual and so correlative to that individual's right not to be harmed (contrasted with a weaker duty to aid); and/or (2) a prohibition against intending harm (contrasted with a weaker duty not to cause or allow harm that is *merely* foreseen).

However, constraints so characterized ignore important moral complexities. Consider the Trolley Case: a runaway trolley will kill five people if a bystander does not divert it onto another track where, he foresees, it will kill one person. Some nonconsequentialists think the bystander may permissibly divert the trolley – killing one to save the five (or even two) – although in other cases they oppose killing one person to save five (Foot 1978). An appropriately complex constraint might better capture nonconsequentialist judgments of cases. If it does, it will capture the precise way in which an individual is thought to be inviolable, that is, protected by a negative right not to be harmed even if the harm would maximize good.

Some say these more precise constraints are absolute; others insist that no constraints are absolute. If constraints may sometimes be permissibly transgressed to produce a sufficiently great good, there will be a threshold on their applicability. This yields what is called "threshold deontology."

Many nonconsequentialists employ a distinctive methodology. They test and develop theories or principles by intuitive judgments on cases. They compare the implications that proposed principles of permissible conduct have for hypothetical cases (such as the Trolley Case) with their judgments about what can be permissibly done in such cases. If the implications of the principles and the judgments conflict, they may develop alternative principles or theories. If implications of principles and judgments are compatible, the nonconsequentialist must still offer a theory identifying the fundamental, morally significant factors that underlie the principles in order for them to be fully justified.

Nonconsequentialism is not merely concerned with prerogatives and constraints, although they have been the focus of contemporary discussion. For example, nonconsequentialists may also propose that there are distinctive ways of aiding people that do not merely try to maximize the good. In the remainder of this essay, I shall explore these points in more detail.

Prerogatives

Moral prerogatives permit an agent to (1) act in ways that do not maximize the impartial good, and (2) act for reasons that stem from his personal perspective, rather than from the perspective of an impartial judge. But how should we make

these prerogatives more precise? Suppose we assign a constant by which each agent can multiply the weight of his personal concerns, so that they can outweigh an impartial good. The result would sometimes conflict with our intuition. For example, no matter how great the noninfinite factor, we can envision some disaster whose aversion would seem to require the agent to sacrifice his most significant projects, even though, intuitively, we do not think he is morally obligated to do so. Yet it seems morally wrong for an agent to multiply his insignificant projects by this same factor so that they often outweigh vital needs of others. It seems more reasonable for the multiplicative factor to depend on the relative importance of the project to the agent, and even to permit the agent to give fundamental projects lexical priority relative to the impartial good. Even this seems an imperfect characterization, since a true prerogative gives the agent the option to care *less* for himself than for others, and this does not seem to be captured by a multiplicative factor greater than one. This is a reason to think that the prerogative represents a concern for one's autonomy rather than for the importance of one's own project from one's own perspective, relative to the interests of others.

Some justify prerogatives by claiming that humans are psychologically predisposed to be most concerned about their own projects. Hence, if they are morally permitted to pursue their nonoptimific projects for personal reasons, they will not be alienated from their fundamental natures (Scheffler 1982). Notice that this justification does not morally limit others from interfering with someone in their quest to maximize the good; it only implies that someone need not always act of his own accord for impartial good for impartial reasons. Second, this nonalienation justification suggests agents should be permitted to try to control what they care about most. However, I should not be able to control someone else's life merely because that is what I care about most. Hence, a theory of prerogatives must specify from an impartial perspective what we are entitled to control from a partial perspective, uncoerced by others. This would connect prerogatives with constraints as part of a theory of individual rights.

Others justify prerogatives by claiming that consequentialist morality is too demanding, for it could require an agent to sacrifice everything to maximize the impartial good. This justification, though, is troublesome since nonconsequentialism can also be very demanding: we may have to make enormous sacrifices in respecting constraints. Why should agents have to sacrifice projects to avoid violating constraints, but not to promote impartial good? To explain this, I believe, we also need a theory of what individuals are entitled to control.

Still others ground prerogatives in the idea that people are ends-in-themselves. Since we should not view people as mere means of promoting the greater good, each of us can sometimes justifiably pursue nonoptimific goals. On this view, the foundation of prerogatives can also be the foundation of constraints on interferences, as both are connected to the idea of personal sovereignty. Even more fundamentally, prerogatives can be seen as a by-product of the fact that moral obligation is not about producing as much good as possible. It is about respect for persons and doing as much good as that requires.

Constraints

Harming versus Not-Aiding

I have suggested that the theory of prerogatives should be connected to a theory of constraints and negative rights. By understanding constraints, we will better understand why we morally must suffer greater losses to avoid violating constraints than to maximize the good. Some nonconsequentialists claim there is a strong moral constraint against harming people who pose no threat and are not responsible for threats, when they do not consent to be harmed. (My discussion here is limited to such people.) Consequentialists argue that there is no intrinsic moral difference between harming and not-aiding (call this the Equivalence Thesis). This thesis implies that we may generally harm to aid. Consequentialists sometimes employ the methodology of intuitive judgments about cases to support the Equivalence Thesis. They identify seemingly *comparable* cases of harming and not-aiding, that is, cases where contextual factors such as intention, foresight, consequences, motive, effort, and so on, are equal. They claim that in such cases we judge that harming and not-aiding are morally equivalent. However, to prove a universal claim like the Equivalence Thesis, one set of comparable cases will not suffice. For it may be that in some equalized contexts, a harming and a not-aiding will be judged equally wrong, yet in other equalized contexts they will not be. What I call the Principle of Contextual Interaction accounts for this possible phenomenon: a property can behave differently in one context than in another. If we can find even one set of comparable cases in which a harming is morally worse than a not-aiding, we rebut a universal claim like the Equivalence Thesis.

James Rachels uses cases like the following Bathtub Cases to prove the Equivalence Thesis: (1) Smith will inherit a fortune if his little cousin dies. One evening while the child is taking his bath, Smith drowns him. (2) Jones will inherit a fortune if his little cousin dies. When Jones enters the bathroom, he sees that the child slipped and fell face down in the water. Although Jones could easily save the child, he does nothing, intending that the child die (1975). Rachels and others following him claim that here a killing and a letting die are morally equivalent and this shows that killing and letting die are morally equivalent per se. But is even the first claim true? Would it be permissible to impose the same losses on Jones and Smith if these losses would bring their victim back to life? I do not think so. Although it might be permissible to kill Smith, it would not be permissible to kill Jones. So perhaps there really is a moral difference between the killing and the letting die, even when they are both morally wrong.

The same point can be made if we ask how much effort an agent must expend to avoid killing someone and to save someone, even in cases where death is equally foreseen or intended. Here is a set of Road Cases: (1) We know that if we drive down one road, we will kill someone who cannot move out of the way. The only

alternative is to go down a side road, where we risk hurting ourselves. (2) We know that to save someone from drowning, we must go down a side road, where we risk hurting ourselves. I think an agent is obligated to face a larger personal risk to avoid killing than to avoid letting die. If this is right, there is a fundamental moral difference between killing and letting die, which are particular types of harming and not-aiding.

These cases suggest killing and letting die are morally different per se; but they do not tell us *why* they differ. We might be able to determine why if we focus on differences that remain in *these cases* after equalizing contexts: (1) in killing we introduce a threat that was not previously present, while in letting die, we do not interfere with a currently present threat; (2) in killing we act, while in letting die we fail to act (in particular, we refrain from acting); (3) in killing we cause someone to lose life that he would have had independently of our actions at that time; in letting die, someone loses only life that he would have had with our help at that time; and (4) in killing we interfere first with the victim, while in letting die, we avoid being interfered with (by having to aid).

These differences might explain the fundamental moral difference between killing and letting die if they are essential (or conceptual) differences between killing and letting die, not just differences in some cases. Are they? Consider (2). Suppose we actively terminate (e.g., pull a plug on) lifesaving assistance we are providing to save Michael from a threat we did not produce, to avoid the substantial effort to us involved in continuing aid. We foresee that Michael will die. In this case (Terminate Aid Case), I believe that we let Michael die even though we *act* to stop the aid (not merely omit to provide it). This letting die is as acceptable as not starting the aid to begin with. Moreover, we are partial *causes* of Michael's death; after all, it resulted because we acted. Hence, we cannot distinguish between killing and letting die simply by saying that the latter involves no action and causes no death.

Still, in the Terminate Aid Case, we do not introduce a cause which induces death. If we did, we would kill. That is, only killings can have this property, although perhaps some killings do not. Moreover, in the letting die case, we stop our being interfered with first and the "victim" loses only what he would have had due to our aid now. I suggest that these are essential properties of letting die but not of killing and hence are essential differences between the two.

We must be careful in speaking of essential properties, for there are two types: (1) those that are essentially true of either killing or letting die per se and also necessarily excluded from cases involving the other; (2) those that are essentially true of one of the dyad, and not necessarily excluded from cases involving the other. The first type creates the most obvious differences; but the latter still creates vital differences even though the properties are "exportable" to an instance of the contrasting behavior. Thus, some *cases* of killing (though not killing per se) could contain what is an essential property only of letting die and vice versa. Nonetheless, these exportable properties could still explain the moral difference between killing and letting die per se. Indeed, rather than compare equalized cases of killing

and letting die, we could compare two cases of killing, alike in all respects except that only the second case has an essential exported property of letting die. As an example, killing someone who is independent of our aid is compared with killing someone who is receiving lifesaving aid from us. If the action in the second case is less morally problematic than the action in the first, then we have strong evidence that this essential property of letting die is morally significant.

If the property functions in the same way on its home ground (i.e., in letting die) and killing has no essential property that can also improve behavior, then letting die would have at least one more morally improving essential property than killing, and hence be morally better per se in virtue of that property. Exportable properties could explain one way we might find cases in which a killing and a letting die were morally equivalent: these could be examples of killing and letting die in which essential properties of one of the behaviors were exported to the case involving the other. Then, as long as other morally relevant properties were equivalent, we would have identified *a* killing and *a* letting die that were morally equivalent. But that would not show that killing and letting die per se were morally equivalent.

I have argued that letting die essentially has properties that make acts morally more permissible. These properties are that the "victim" loses only life that he would have had with the agent's help at that time and that the agent prevents his being imposed on first. But these properties can be morally important only if we have a stronger claim than others do to what we have independently of the current aid of others when the aid does not merely counteract their initial threat, and this stronger claim applies both to our life and to the efforts we could make on behalf of others. The moral distinction between killing and letting die captures this view of separate persons.

We must exert great effort to avoid imposing first on others, especially on what they have independently of our current aid. We can legitimately make fewer efforts to prevent someone from not having what he would have only by our being imposed on first to provide him aid. We can now see that making great efforts not to kill others, at least when the case does not share certain essential properties of letting die, is consistent with a prerogative not to maximize the good by aiding others. However, if we explain the moral distinction between killing and letting die as I have, we must do more work to explain why killing in the Trolley Case to save five is permissible.

Finally, there are certain things we should remember in applying our conclusions about killing and letting die to the general moral distinction between harming and not-aiding. (1) When we kill or let someone die, we might reasonably think that she has some right to her life. But when we harm or do not aid someone in non-life-and-death cases, what they lose – or fail to get – may be something to which they have no right. (2) Generally, when someone kills, they interfere with another's body in ways they do not do when they let die. However, if we harm someone in non-life-and-death cases, we may not necessarily interfere with her body any more than if we do not aid. Suppose we combine these two factors and

construct a set of harming and not-aiding cases: (i) some money that does not belong to anyone is accidentally transferred to my bank account, but you make me worse off by transferring it out; (ii) you fail to transfer some unowned money into my account. There may be no great moral difference between these cases, though in the first, someone is made worse off while in the second we fail to improve his condition.

Intending versus Foreseeing Harm

The Doctrine of Double Effect (DDE) is historically the most important formulation of the supposed moral distinction between intending and foreseeing harm. The doctrine states that there is a moral constraint on intending evil, even when the evil will be a means to a greater good. Nonetheless, we may be permitted to employ neutral or good means to pursue a greater good, even though we foresee evil side effects if (1) the good is proportionate to the evil and (2) there is no better way to achieve this good. Thus, it is said to be impermissible to end a war by intentionally killing ten nonthreatening civilians (Terror Bombing Case), but permissible to end the war by intentionally bombing munitions factories, even foreseeing that twenty civilians will certainly die as an unintended side effect (Tactical Bombing Case).

The supposed moral distinction between intending and foreseeing bad effects applies to omissions as well as to actions and is independent of the harming/not-aiding distinction. Some nonconsequentialists embrace only one of these distinctions; others embrace both. Moreover, some nonconsequentialists wish to revise the DDE so that it is a nonabsolute constraint. For example, they revise it so that it does not apply at all in situations of self-defense against threatening people, and in other situations it only implies that we must tolerate worse consequences before acting intending bad effects than we have to tolerate before acting with merely foreseen bad effects.

Many object to the DDE because we can typically describe behavior it supposedly rules out so that the agent does not strictly intend any evil. For example, the Terror Bomber might act intending only that the civilians appear dead until peace is declared. Of course, he foresees with certainty that civilians will die since the only way to make them appear dead until the war ends also leads to their death. But the Tactical Bomber also foresees with certainty the deaths of civilians (Bennett 1981). We might try to recapture the moral distinction between these two cases by revising the DDE. The revised version would prohibit intending even minor intrusions on, or involvement of (Quinn 1993), persons when the agent foresees that they will suffer significant harm to which they did not consent. This is a significant revision to the doctrine. The original DDE barred agents from aiming at evil as a means. The revision prohibits agents from intentionally treating persons as tools whenever the results would be foreseeably bad for them, even though there is no intention that bads occur as a means.

The DDE is also problematic because it never rules out producing a greater good by necessary means merely because of the lesser bad side effects of the means. As Philippa Foot notes, it permits us to use a gas to save five people even knowing the gas will seep next door killing one person. It would also permit us to rush to the hospital to save five, foreseeing (but not intending) that we thereby run over and kill one (1978, 1984). Yet, Foot claims, intuitively we think it is impermissible to do these things.

The traditional DDE is also too strong. It seems to rule out intentionally harming someone to promote that person's overall good (the *intra*personal case), and it rules out intentionally harming someone to help others even when that person is no worse off than he would have been otherwise.

Two further complexities: The DDE suggests that the *greater* good against which the bad side effect is compared must be *intended*. But can it not be a mere foreseen side effect of what was intended? For example, in the Massacre Case a strategic bomber plans to destroy one portion of a munitions factory. He intends to bring about this small good but foresees two side effects: (1) killing ten innocent civilians; and (2) stopping a massacre of twenty different civilians. (1) is too large an evil to be outweighed by the small intended good. (2) is a great enough good to outweigh (1), but it is not intended since its occurrence is not necessary to the war effort. Hence, if the DDE is a necessary condition for moral permissibility, it might block the attack on the factory though it is permissible.

Notice that if the bomber in this case proceeds only *because* (2) will occur, this need not imply that he intended to produce that good. This suggests that what is known as the Counterfactual Test for detecting intention (rather than mere foresight) is flawed. That test states that if we would not proceed with our act had a particular effect *not* occurred – assuming everything else is held constant – then in acting we intend that effect, as a means or as an end. However, as the previous example suggests, in some cases we might proceed only because an effect will occur, yet still not intend its occurrence. This distinction between doing something *because* an effect will occur and doing it *in order that* it will occur suggests that there is a third type of case between Tactical and Terror Bomber: Suppose it is militarily valuable to bomb a munitions factory only if it is not immediately rebuilt. The factory will be rebuilt unless the population is grieving as a consequence of the death of civilians in the bombing. Hence, we bomb the factory only if we foresee with certainty that civilians will die, even though we do not intend that they die (Munitions Grief Case). I believe it is permissible to bomb in this case, even if terror bombing is impermissible. Because of this third type of relation to effects – because they will occur – it might be better to speak of the *Doctrine of Triple Effect*.

Finally, a major objection raised to the DDE (Thomson 1999; Scanlon 2008) is that states of mind (such as intending) cannot make an act that is otherwise permissible be impermissible. For example, imagine a variant on Tactical Bombing in which everything is the same (the munitions are bombed and this causes a proportional number of civilians to die) except the bombardier's intention in

bombing the munitions are to cause the civilians' deaths. His intention does not seem to affect the permissibility of bombing, which depends on the necessity of destroying the munitions and the proportionate loss to civilians.

Complications on the Simple Constraints

As I noted earlier, many contemporary nonconsequentialists want to develop W.D. Ross's conception of prima facie duties. Ross thought that when such duties conflict, we have no rule or principle ranking them. So some nonconsequentialists have tried to develop more complex and less frequently overridable duties. For example, the Trolley Case suggests that we might need to more precisely characterize the duty not to harm so that it does not conflict as often with a duty (or desire) to aid. We are looking for a principle that explains why it is permissible for a bystander and/or for someone driving the trolley to help several people by redirecting a fatal threat so that it kills someone else, yet it would be impermissible to kill one person to harvest his organs to save others (Transplant Case). The principle must also explain why some things we could do to stop the trolley (e.g., pushing an innocent heavy onlooker into its path) are as impermissible as harvesting someone's organs, whether done by the trolley driver who is trying to avoid killing even more people or by a bystander who is merely trying to help.

Philosophers have offered many ways of explaining these intuitive judgments. Among them are: (1) When (i) we redirect the trolley, we merely foresee the death of the one; when (ii) we harvest the organs for transplant, we intend the death; and when (iii) we push the bystander into the trolley, we intend his involvement and foresee his death. Hence, (i) is permissible and (ii) and (iii) are not. However, this DDE-inspired explanation suggests that a trolley driver or bystander could legitimately detonate a bomb to stop the trolley, even though they foresee, but do not intend, that the bomb will involve and kill a bystander. However, I believe this is impermissible. (2) In Trolley, we do not initiate a new threat. We merely redistribute a preexisting threat so that a greater number are saved. But this, even in combination with (1), cannot be a sufficient condition for acting permissibly. If a trolley is headed toward one person, we may not redirect it foreseeing it will kill five, even if we do this because the redirection also moves a rock that saves twenty people from another threat. Neither is (2) a necessary condition for acting permissibly. Suppose a trolley is headed toward five people seated on a large swivel table. Although we physically cannot redirect the trolley, we can turn the table and save the five. However, we thereby start a rock slide which will kill one innocent bystander (Lazy Susan Case). Here we start a new threat which kills someone. Nonetheless, I believe it is permissible to act. The problem is explaining why.

One proposal is the Principle of Permissible Harm (PPH) (Kamm 1989, 1996, 2007). The basic idea is that it is permissible for (1) greater good and (2) means

that have greater good as their noncausal flip side to cause lesser evil, but not permissible for an act (3) to require lesser evil (or someone's involvement leading to lesser evil) as a means to greater good, or (4) to directly cause lesser evil as a side effect when the act has greater good as a mere causal effect. "Noncausal flip side" signifies that the greater good occurring is, in essence, another way of describing the situation in which the means occur. This principle denies that we may never cause serious harm to people who need not volunteer for such harm, in order to aid others. For example, when harm is an effect of achieving a greater good, we may permissibly do what harms. Suppose by directing gas into a room we can save five people. However, their breathing normally – the greater good – alters the air flow in the room, redirecting germs and killing an innocent person. In this case, it is permissible to use the gas to save five people, because it is the greater good itself which causes the death.

The PPH explains why we may permissibly turn the trolley. The trolley moving away, which kills the one, is a means to saving the five and this greater good is its *noncausal* flip side. That is, given that the moving away occurs in a context where no other fatal threat faces the five, the five's being saved just is the trolley moving away. Further, our act of turning the trolley, which ultimately leads to harm, is permissible by (2) because it produces the harm only by producing a means (the moving trolley) that has greater good as its noncausal flip side. By contrast, a bomb that kills a bystander saves the five as a mere causal effect of its moving the trolley away from them, so the act that sets the bomb off is impermissible according to the PPH.

A problem for the PPH is the Loop Case in which a trolley is headed toward the five and it can be redirected onto another track where one person sits; however, the track loops back towards the five (Thomson 1985). Either the trolley will kill the five from its original direction or, if we redirect it, it would kill the five after it loops were it not that it hits into the one person and is stopped. I believe it is permissible to turn the trolley in this case. Yet, hitting the one is a required causal link to saving the five; it is not merely a foreseen side effect. Does this mean that if we turn the trolley, we intend the hitting of the one? Presumably we would refuse to turn the trolley unless this happened to the person, for if the trolley did not hit him, five people would die anyway and we would also harm him on its way to the five. In short, we turn the trolley *because* he will be hit. But, as I noted earlier in discussing the Counterfactual Test, that need not imply that we intend to hit him. Consequently, our judgment in the Loop Case is consistent with a revised version of DDE being a necessary condition of permissibility. It also shows that a rational agent can pursue a goal that he knows is achievable only by a certain causal route without intending that route.

However, the Loop Case may still be a problem for the PPH. For if hitting the one is causally necessary to save the five, how can the greater good, or means which have greater good as their noncausal flip side, produce the lesser evil? We might revise the PPH as follows: when the trolley heads away from the five, we are left with a *structurally equivalent component* of the greater good – that is, what

would be the greater good if it could be sustained. This is so because the only threat the five still face – the trolley coming at them from another direction – arises only because we removed the initial threat. The structural equivalent of the greater good, or means that have it as its noncausal flipside, produces a new problem as well as the means for eliminating it (the hitting of the person), and this makes turning the trolley permissible. So the PPH should be revised to allow that a structural equivalent of the greater good or means that have it as a noncausal flipside may produce lesser evil, even when this lesser evil is causally necessary to *sustain* the greater good (even if not to produce it).

What morally significant ideas might justify the PPH? I have suggested that it gives expression to a moral distinction between substitution and subordination of persons (Kamm 2007). That is, it is sometimes permissible to *substitute* a person for others so that the substituted person is threatened instead of the original people. This contrasts with *subordinating* one person to others, as occurs in the Transplant Case. However, such substitution that leads to harm is permissible only when it is caused in certain ways – for example, as a result of greater good causing the lesser evil.

Inviolability

The PPH (or principles like it) implies that persons have rights not to be treated in certain ways simply to save more lives. These rights protect persons against some ways of maximizing the good: it gives them some inviolability. The inviolability is not absolute. It is limited *qualitatively*. (That is, the PPH permits some ways of harming.) It may also be limited *quantitatively*. (For example, the PPH might be overridden to save a million people.) The former limitation accords with the PPH; the latter is an external restriction on it.

Another way in which the PPH may be limited is by what I call the *Principle of Secondary Permissibility* (PSP). For example, in the first instance it is impermissible to push an innocent onlooker into a trolley that will amputate his leg in order to save five other people from the trolley's killing them. However, suppose the alternative is to redirect the trolley away from the five and toward that very same person, thereby killing him. Redirecting it toward him would ordinarily be permissible and is, suppose, something we would do if we could do nothing else. However, since it is in his interest to lose his leg rather than be killed and both are against his will, *secondarily* it becomes permissible, I think, to push him into the trolley, an action that was not, in the first instance, permissible. Indeed, this may become the *only* permissible harmful act.

Are people so inviolable that agents may also not violate PPH restrictions on harming one person, even if that is the only way of minimizing violations in others of the PPH itself? The claim that we may not violate someone's rights to minimize violations of others' comparable rights is sometimes called the "paradox of

deontology." Some claim that if we really *care* about rights, we should minimize their violation even if this requires us to violate comparable rights. Those who agree with this say they cannot see how one person's right could stand in the way of minimizing the violation of comparable rights. If they nevertheless think we should not violate the restrictions of the PPH, it is because they are concerned with the agent who would act, not with the rights of the potential victim per se. This model derives the constraints on violating rights to minimize rights violations from "inside (the agent) out (to the victim)" rather than from "outside the agent (in the victim's right) in (to the agent)" (Anderson 1993; Darwall 1986). Does such an agent-focused approach explain the constraint?

The agent-focused explanation of the constraint on minimizing rights violations has frequently employed the idea of agent-relativity. On this view, each of us has duties that are fundamentally relative to the particular agent we are. Some argue that both consequentialist and nonconsequentialist theories can embrace agent-relativity. For example, some consequentialists say that, although each agent has the same agent-neutral duty to produce the best state of affairs, from each agent's perspective the state of affairs in which he kills one person is worse than one in which another agent kills more people (Sen 1982). Hence, each individual has a duty to avoid *his* killing someone even to reduce a greater number of killings by others.

This is an agent-relative consequential system since there are multiple agent-relative best outcomes, not just one agent-neutral best outcome that different people are in different positions to forward. But how can this approach explain a constraint (which I believe does exist) on my killing one person in order to save a greater number of people whose rights I *myself* either have endangered or will endanger? If I do not kill the single person, the consequence will be a world in which I am the killer of a greater number of people, and this seems like the worse world from my perspective. So if, according to this approach, I must produce the best world, I should kill the one. This, I believe, is the wrong conclusion.

A nonconsequentialist agent-relativist might argue that we have special responsibilities to our victim (who is the person we will kill, not the ones we let die), even if killing him would promote better agent-neutral consequences. That is, our victim's interests are magnified from our perspective (Fried 1978; Nagel 1986). However, if the only way to save a greater number whose rights we ourselves have endangered, or will endanger, is by killing the one, why should our special responsibility to our many victims not dictate that we kill the one? Yet, this is the wrong conclusion.

In order to avoid these problems, both consequentialist and nonconsequentialist agent-relativists might give special weight to an agent's present acts. They might claim that we are especially responsible for what we do and what we produce *now*, by contrast with our past and future acts. But why should our current actions and their consequences take moral precedence over our past or future ones? Why should *we now* be so important?

There are, I believe, agent-focused views that are not essentially agent-relative. While they focus on the quality of an agent's act or state of mind, rather than on a victim's right, they do not take note of the "agent's mark" on the act, victim, or outcome. For example, the quality of the act or state of mind in which an agent must engage if he kills the one person is found repellant. The act would be the agent's if he did it, but it is not essentially its being *his* rather than what it is in itself that repels him (Nagel 1986; Williams 1981). Advocates of this view might claim it explains why someone should not kill one person to save a greater number of people even from her *own* future bad acts. The explanatory structure of this duty-based constraint is essentially the same as a rights-based constraint. In both, one instance of either an act type or right type stands in the way of minimizing misconduct involving many instances of the same act type or right type. However, if the logic of concern for this duty does not require that we minimize instances of its violation, but simply not violate it, why does the logic of the concern for a victim's right require that we minimize violation of the right?

Now consider the Art Works Case: If someone loves beauty, he will be disposed to preserve and not destroy art works. What should this person do if he must destroy one artwork to preserve several equally good ones? Presumably it is permissible for him to destroy one to save the five even though the act of destruction is repellant. This suggests that the constraint on harming *persons* is not derived from inside the agent out, but from *outside* her in, since the constraint reflects the kind of entity she would act on – a person, not a work of art.

Consequently, I advocate a victim-focused, rights-based account of constraints. Are there any problems this approach cannot explain? Suppose the only way we can prevent five people from being killed in violation of the PPH is to kill one person, A, in violation of the PPH. Does it make sense to express concern for the inviolability of the five by treating A as *violable* for their sakes? But then morality would say that sometimes it is permissible to treat people inconsistently with PPH restrictions, and this just means that people are less inviolable than they would be if it were impermissible to do this. It is true that if we do not kill A, more people will be seriously violated. But this does not mean that their *inviolability* is less. Inviolability is a status. It defines what we can permissibly do to people rather than what actually happens to them. If the five are killed because A is not killed, morality does not endorse (make permissible) their being killed. By contrast, if it were permissible to kill A to save the five, the *inviolability of all* six would be lower. After all, to permit the killing of A implies we may kill anyone else in similar circumstances and that morality *endorses* killing people in this way.

The explanation I have offered for why it is impermissible to kill A to save others from being killed puts emphasis on what it is permissible to do to people rather than on what happens to them. Unlike the agent-relative account, it does not focus on what I do rather than what others do. The fact that if I kill someone, I would be acting now and the victim would be mine, does not play a pivotal role in explaining why I must not kill him. We explain that by focusing on each person's inviolability. His right, not my agency, constitutes the moral constraint. The fact

that the other five have this same right does not diminish the constraint against violating the one's rights that I come up against.

Thus, my account highlights an *agent-neutral value*: the high inviolability of persons. Each agent must respect this value and does so in being constrained by the rights of the first person he encounters, even though the identity of this person will differ for each agent. This agent-neutral value is not a consequentialist value we bring about through action or omission. The value already resides *in* persons.

If a person has a high degree of inviolability, she will have a strong right protecting her. We will owe it to her not to transgress the right, and will wrong her if we do. Hence, another way to put the argument I have given for not killing the one is that the importance of persons can be expressed by rights being strong, rather than by their being weaker so that we may minimize violations of them by transgressing them. It would be *self-defeating* for it to be permissible to violate a strong right, which itself claims that someone should not be used in order to stop rights violations, in order to stop comparable rights violations.

If people are morally inviolable in a certain way, then, I believe, they have a *higher* – and not merely a different – status. It might be argued that persons have higher status if saving them is so important that we must harm one of them to prevent harm to many. But we must remember that if the one person may be sacrificed, then those others may, in the appropriate circumstances, also be sacrificed, and this lowers their status. Most importantly, if saving only a *greater number of people* could make the killing of one permissible, the importance of saving people would not reveal anything about the status of any *individual* person.

Suppose people have a right not to be harmed even to minimize violation of comparable rights of *others*. From behind a veil of ignorance (the ex ante perspective), no one knows whether she would be the single person sacrificed or one of the many whose rights would be protected. However, everyone would know that her chances of being one of the many who would be saved are greater than of being the one sacrificed. Why would it not be rational for each to agree to forego a right not to be sacrificed for others when, after all, this would reduce the chances that one's own right would be violated?

Moral theories seeking to maximize each person's ex ante probability of some good would justify killing in many cases. Suppose members of a community consider purchasing an ambulance. They know they will save more lives if they have one, but they also foresee that in speeding through town, the ambulance will kill a few people. Now, imagine that we can save still more lives by attaching a device to the ambulance which prevents the driver from swerving to miss a pedestrian whenever swerving would decrease the number of people who live. Using this device would maximize the ex ante probability of survival of each person (Ambulance Case). Nonetheless, I believe an agreement to use the device would not make its use legitimate. In general, we cannot permissibly "bargain away" our moral status not to be treated in certain ways in order to increase our life prospects or to minimize rights violations. Our moral status is in this way inalienable. (Although

we may permissibly waive our rights supererogatorily and voluntarily sacrifice ourselves when we want to save others.)

Yet it is important to delineate when it is and is not true that we cannot bargain away certain inviolabilities. For consider the Two Diseases Case (Kamm 1996), in which there are two diseases in a community. One, the Arm Disease, causes one and only one arm to fall off, and is very prevalent among a part of the population whose members we can identify beforehand. The second, the Death Disease, is very rare in a *different* part of the population that we can identify as susceptible to it. The only thing that cures the Arm Disease is a serum made from the finger of a person who was subject to the Death Disease but did not get it, and the only cure for the Death Disease is a serum made from the arm of a person subject to the Arm Disease who did not get it.

I believe it would be in the interest, ex ante, of all involved to make an agreement to provide the resources necessary to make the serums needed at the time they are needed, and that enforcement of this agreement ex post (i.e., once one knows who will and who will not be getting the diseases) would *not* be morally wrong. There is a high incidence of the Arm Disease; so there is a high probability that the people once susceptible to the Death Disease will lose a finger in exchange for avoiding the small risk of a big loss to themselves, that is, death. There is a low incidence of the Death Disease; so there is a low probability that a person once susceptible to the Arm Disease will lose an arm – that is, a low probability of his making a larger sacrifice for others (than those susceptible to death make) in order to diminish his chances of losing an arm (i.e., in order to lower a high probability of his suffering the loss of an arm).

This is a case in which an Arm person would have to pay with the *very item* he had attempted to increase his probability of keeping (his arm), at a time when it is known to be no longer in his interest to do so. The significant difference between this case and cases considered before, however, is that here the arm would be sacrificed to prevent an even greater loss (death) to another person. What the person who is sacrificed loses is significantly less than what the person who is saved would lose if he were not saved. Additionally, there is potential reciprocity of sacrifice. The Two Diseases Case helps us see that an agreement to have an arm taken is acceptable *not* merely because the person does not lose the very thing he was trying to insure. This example presents us with a situation in which an actual self-interested agreement could legitimize losses which could not be legitimized simply by the goal of maximizing lives saved.

Nonabsoluteness of Constraints

Even internally complex constraints such as the PPH might not be absolute. Although nonconsequentialists must explain when they may be overridden, I shall

not attempt to do that here. The point I wish to emphasize is that even if the constraints might be legitimately overridden to achieve some greater good, this need not imply that they may permissibly be overridden in pursuit of personal goals – not even if the pursuit of those same personal goals legitimizes failing to pursue that greater good. The relationship here seems intransitive. Suppose "G" stands for "greater good"; "P" for "personal interests and goals"; "C" for "duty to respect a constraint"; and ">" means "may permissibly override." (P > G) and (G > C) may both be true, and yet (P > C) may not be. Suppose someone insisted on transitivity. Then she would need to deny that (P > G) (i.e., deny prerogatives) or hold that constraints are absolute in order to avoid P > C.

To defend the intransitivity thesis, we shall assume our discussion of prerogatives explains why P > G and try to show that sometimes G > C. Ordinarily, promises morally constrain us. Yet it might sometimes be permissible to break even an important promise (e.g., a bodyguard's promise to protect her employer's life) to save thousands of people. We might permissibly break the promise even if saving the thousands is supererogatory because the sacrifice required of us to save them is great. This supports the claim that G > C even if P > G. Nevertheless, we might be required to suffer grave personal loss to respect the constraint (e.g., the bodyguard might have to endanger her life to keep her promise). Hence, –(P > C). We now see that there are two ways to measure the moral significance of acts: (1) how great a personal loss we are required to suffer to perform them; and (2) the capacity of one act to take precedence over another. Maximizing the good may be more important by measure (2), but not by measure (1); abiding by constraints may be more important by measure (1) and not by (2).

How can we explain the apparent intransitivity and the conflicts between these two measures? Constraints are minimum standards we must all meet. We may be required to sacrifice our personal goals to meet these standards, but not to go beyond them. That explains P > G even if –(P > C). Someone might suggest that G > C if the loss *to* the agent of not achieving G exceeds the amount she would have to sacrifice to respect C. But someone might violate C for G even though she cares more about C than G. The evidence for this is that she would suffer a greater personal loss to abide by C than to bring about G. In short, the proper solution is not to "personalize" the loss of G. Rather, the agent understands that promoting the greater good is, from an impartial perspective, morally more important than respecting the constraint.

In essence, my account explains the intransitivity in the relation among prerogatives, constraints, and the pursuit of the greater good in a nonconsequentialist theory by noting that the precedence relation in each premise is based on a different factor: P > G reflects the entitlement of each individual as an end-in-herself not to sacrifice for the greater good; G > C reflects the impartial weight of the good; while –(P > C) reflects the moral importance of minimal standards in relation to personal interests. We should not expect transitivity if different factors account for precedence relations. (Even if the same factors explained the

precedence relation in the first two premises, intransitivity may still arise from what I call the Principle of Contextual Interaction: the interaction of P and C could produce a new factor not present when P and G and G and C interact.)

Nonconsequentialist Principles for Aiding and Aggregating

Nonconsequentialism not only tells us when there is a duty to aid, but very likely offers distinctive principles of *how* to aid that may conflict with the goal of maximizing the good. It also may provide distinctive reasons for doing what maximizes the good. In this section, I shall consider these principles and reasons.

Suppose we cannot help everyone in need because each needs some scarce resource. Different principles exist for different situations: (1) there may be true scarcity such that more of the resource will not appear; (2) there may be temporary scarcity, so we can eventually help everyone; (3) we may be uncertain whether we are in (1) or (2). I shall focus on (1).

Suppose we are dealing with two-way conflict cases between potential recipients. When there are an equal number of people in conflict who stand to lose the same if not aided and gain the same if aided (and all other morally relevant factors are the same), fairness dictates giving each side an equal chance for the resource by using a random decision procedure. But there may be a conflict situation in which *different* numbers of relevantly similar people are on either side and they stand to lose and gain the same thing. This raises the question of whether nonconsequentialism requires us to give each person an equal chance to be helped, or permits us to aggregate and help the greater number of people.

Some have argued that in conflicts like this, it is worse for the greater number if they die but better for the lesser number, and there is no impartial point of view from which to judge that it is worse if more die (Taurek 1977). However, the following Argument for Best Outcomes suggests this view is flawed (Kamm 1993): (1) Using Pareto Optimality, we see that it is worse for both B and C to die than for only B to die – even though it is not worse for B. (2) It is worse to a still greater degree if B, C, and D die. Our judgment that the world is worse to a greater degree, although it is also only worse for one additional person, by comparison to what is true if B and C die, is made from a point of view outside that of any person (this goes beyond Pareto Optimality). (3) A world in which A dies and B survives is just as bad as a world in which B dies and A survives. This is true from an impartial point of view, even though the worlds are not equally preferred by A and B. (4) Given (3), we can substitute A for B on one side of the moral equation in (1) and get that it is worse if B and C die than if A dies. Hence, nonconsequentialists, as well as consequentialists, can evaluate states of affairs from an impartial point of view.

Although it would be worse that B and C die than that A dies, that does not necessarily mean it is right for us to save B and C rather than A. As

nonconsequentialists, we cannot automatically assume it is morally permissible to maximize the good, for this may violate justice or fairness. Some might claim that if we save B and C on the basis of (4), we abandon A to save the greater number without giving her a chance and this is unfair. They might object that we could generate an intransitivity, where ">" means "clearly ought to be saved," B + C > B quite strictly, A = B quite strictly, but –(B + C > A), because it could be unfair to deprive someone of his chance.

But is it really wrong to produce the best outcome in this case? Here are two arguments against its being wrong. The Consistency Argument indirectly shows that in saving the greater number, we need not be overriding fairness or justice: in many other cases, nonconsequentialists will not violate justice to save the greater number. For example, they often will not kill one to save five. Moreover, they (arguably) would not deprive a janitor of a chance for an organ transplant simply because a doctor who can save the life of a third party also needs the organ. Why would nonconsequentialists refuse to sacrifice justice or fairness to save more lives in these cases but override justice or fairness to save even two lives rather than one in simple allocation cases? It is most reasonable to believe they choose in this way because fairness is not being overridden. If so, fairness does not require that we give A a chance.

Second, the Balancing Argument (Kamm 1985, 1993) claims that in a conflict, justice demands that each person on one side should have her interests balanced against those of one person on the opposing side; those that are not balanced out in the larger group help determine that the larger group should be saved. If we instead toss a coin between one person and any number on the other side, giving each person an equal chance, we would behave no differently than if it were a contest between one and one. If the presence of each additional person would make no difference, when this affects their good, this seems to deny the equal significance of each person. Thus, justice does not conflict with producing the best outcome. (Some might suggest we should give chances in proportion to the numbers of people in each group, but I think this is a mistake.) Hence, aggregation might be required, but for distinctly nonconsequentialist reasons.

How might we extend the nonconsequentialist principles to conflicts when the individuals are not equally needy? Consider a case where the interests of two people conflict with the interests of one. The potential loss and gain of the one is equal to the potential loss and gain of one of the other two. The potential loss and gain of the second of the pair is less than those of the others. A consequentialist claims we must maximize good and, therefore, choose to help the pair. A contractarian arguing behind the veil of ignorance might agree if she is trying to maximize the ex ante expected good of each person. Must a nonconsequentialist, committed to balancing equals, do the same? No – at least not always. Suppose the lesser loss is a sore throat and the greater loss is death (rather than living for ten more years). Call this the Sore Throat Case. To preclude someone's chance to live in order to gain a sore throat cure is to fail to show adequate respect for a person's life, since from her partial point of view she is not indifferent between her survival and the

survival of one of the pair. In short, although helping the pair is morally better than helping only one of them, and helping one dying person is morally as good as helping another, helping the pair is not necessarily morally better than helping the single person in this case. (Notice that the ground for ignoring the small extra good is not simply that we should not think of such matters in life and death situations. For it would not be wrong to choose between two decision procedures that give each of two people an equal chance on the ground that one procedure will also magically cure someone's sore throat.)

This form of reasoning gives equal consideration to each individual's partial point of view from an impartial point of view, so it combines subjective and objective perspectives. Hence, I call it *Sobjectivity*. It implies that certain extra goods (like the throat cure in its role in the Sore Throat Case) can be morally irrelevant; I call this the Principle of Irrelevant Goods. Whether a good is irrelevant is context dependent. Curing a sore throat is morally irrelevant when others' lives are at stake, but not when others' ear aches are. The Sore Throat Case shows we must refine the claim that what we owe each person is to balance her interests against the equal interests of an opposing person and let the remainder help determine the outcome.

We might explain this conclusion by saying that any loss or gain (X) that is significantly less than N, and so could not be a contestant on its own against N, cannot legitimately determine any distribution in combination with N losses or gains. I call this Sobjectivity 1. But suppose X is saving someone's leg? We should save one person's life rather than someone else's leg when these are the only morally relevant considerations. Perhaps, though, it is better to save one person's life and a second person's leg than to give a third person an equal chance at having his life saved. If so, we might embrace Sobjectivity 2, which is based on the following reasoning: According to the nonconsequentialist, each of us has a duty to suffer at least a relatively minimal loss (e.g., a sore throat) to save another person's life, and if it matters to each person that his be the life saved, we each have a duty to suffer a minimal loss to give someone else a chance at life. Further, so long as suffering the small loss is a duty for any given person, *no number of the smaller losses can be aggregated* with another's life to outweigh someone else's chance to live. Where the loss X is greater than the required loss (e.g., losing a leg), then we should prevent N + X rather than give someone an equal chance to avoid N. By contrast, according to a consequentialist, what an individual has a duty to do has nothing to do with what may or may not be aggregated, and an aggregate of small losses can outweigh a greater individual loss.

How about the following cases? Suppose, according to a nonconsequentialist, that no one has a duty to lose three fingers to save a life. From an impartial point of view, we might still think that giving one person a chance at life is more important than saving another's three fingers when combined with saving a third person's life. If so, we should reject Sobjectivity 2 for Sobjectivity 3, which insists that it is from a view outside that of any of the party's duties that we decide matters. Hence, further characterization of the relevant and irrelevant goods might

be as follows: (a) a certain good would be relevant in a choice between two lives, in the sense of making the side to which it is added deserve a greater proportional chance of getting aid, if the aggregation of many instances of it alone could have proportional weight against saving a life; (b) an extra good could be determinative of our choice when it is conjoined with one life against another life (i) if that good on its own merits a proportional chance against a life when the choice is about whom to aid, or (ii) an aggregation of many instances of it alone could directly outweigh saving a life.

It is possible that we should employ one form of Sobjectivity to choose whom to aid "here and now" (e.g., in an emergency room) and adopt another form to make *macro* decisions (e.g., whether to invest in research to cure a disease that will kill a few people or in research to cure a disease that will only wither an arm in many). For example, Sobjectivity 4 for macro decisions might permit aggregation of significant (not insignificant) losses to many people to outweigh even greater losses to a few, even when no individual person in the larger group will lose as much as each individual in the smaller group will lose. Thus, Sobjectivity 4 (and Sobjectivity (3) [bii]) is in conflict with common components of contractualist theories. These are helping the worse off first and pairwise comparison (which requires that the side we help must have at least as many people who will be as badly off individually as those on the other side).

Sobjectivity 4 need not imply that many lesser losses (e.g., of one arm in many people) are the *equivalent of a* life they can outweigh, in the way that one life is the equivalent of another life. Rather, Sobjectivity 4 implies that we will not bear the *cost of* many arms to save a life. This is supported by the fact that Sobjectivity 4 (unlike earlier forms of Sobjectivity) should not be used to decide whom to *harm* in order to aid others. For example, if a threat were headed toward any number of people who would each lose an arm, it would be wrong to turn the threat toward the one person who would be killed. This contrasts with the permissibility of turning a threat away from two people who would be killed and toward a different person who will be killed.

Could Sobjectivity 4 be defended by arguing that just as it would be rational for each individual to bear a small risk of death (e.g., from taking a medicine) in order to have a good quality of life, so when there is a high probability of losing an arm (since many people will lose arms), each may accept a low probability of dying without aid (since only a few will die) in order to save arms? In the multi-person, but not the single-person, scenario, we know that someone will die. The question is whether this is a morally significant difference. Finally, suppose that we did argue for the permissibility of investing in cures for truly minor problems affecting many, such as headaches, rather than in a cure for a rare fatal disease, on the ground that it is reasonable for each person to take a small risk of being the one who will die in order to have headache cures at hand for his many, certain-to-occur headaches. This does not imply that here and now we should not save someone from dying from the rare fatal disease, if we can, rather than cure millions of headaches. For example, suppose that giving all of the aspirin that has been

produced to cure headaches to someone who develops the fatal disease could, surprisingly, save him. It could be wrong to leave him to die on the grounds that it was reasonable ex ante for each person to take a small risk of dying if he fell fatally ill in the future (due to our not investing in an expected cure). It is here and now, when the probability for a particular individual of dying is high, that the irrelevant utilities of headache cures do not aggregate to override saving the life.

A nonconsequentialist theory of the distribution of scarce resources should also consider whether certain characteristics that one candidate has to a greater degree than another are morally relevant to deciding who gets the resource. I call this the problem of interpersonal allocation when there is *intrapersonal* aggregation, because one candidate has all the characteristics the other has and more. Principles described above that apply when the additional goods on one side are distributed over several people may have to be revised so as to apply when additional goods are concentrated in one person rather than another.

A system I suggest for evaluating candidates for a resource starts off with only three factors – need, urgency, and outcome – but it could add other factors later. Need is here defined as how badly someone's life *will have gone* as a whole if he is not helped. Urgency is defined (atypically) as how badly someone's life *will go* if he is not helped. Outcome is defined as the difference in expected outcome produced by the resource relative to the expected outcome if someone is not helped.

The neediest may not be the most urgent. Suppose A will die in a month at age 65 unless helped now and B will die in a year at age 20 unless helped now. I suggest that B is less urgent but needier, since one's life will have gone worse (other things equal) if one dies at 20 rather than at 65. To consider how much weight to give to need, we hold the two other factors constant and imagine two candidates who differ only in neediness. A consequentialist argument for taking differential need into account in cases where life is at stake could be that there is something like diminishing marginal utility of life (i.e., a better outcome is provided if we give a unit of life to those who have had less). I do not think this is necessarily true.

One nonconsequentialist argument for taking differential need into account is fairness: give to those who, if not helped, will have had less of the good (e.g., life) that our resource can provide before giving to those who will have had more even if they are *not* helped. Fairness is a value that depends on comparisons between people. But even if we do not compare candidates, it can simply be of greater moral value to give a certain unit of life to a person who has had less of life (McKerlie 1997).

But need will matter more the more absolutely and comparatively needy a candidate is, and some differences in need may be governed by a Principle of Irrelevant Need. This is especially so when each candidate is absolutely needy; a big gain for each is at stake, and if the needier person is helped he will wind up having more of the good (e.g., a longer life) than the person who was originally less needy than he.

Suppose there is conflict between helping the neediest and helping the most urgent (where outcomes are the same). I claim that when there is true scarcity, it can be more important to help the neediest than the most urgent, but if scarcity is only temporary, the urgent should be helped first, since the neediest will be helped eventually anyway.

Still, there are constraints on the relevance of need in a nonconsequentialist theory of distribution. Giving a resource to the person who will have had less overall of the good it can provide may be impermissible if it fails to respect the rights of each person. For example, consider another context: If two people have a human right to free speech, how long someone's right has already been respected may be irrelevant in deciding whom to help retain free speech. If having health or life for a number of years were a human right, it might not be appropriate to allocate resources on the basis of the degree to which people's rights have already been met.

Now we come to outcome. A consequentialist might consider all effects of a resource. I suggest that for nonconsequentialists (1) effects on third parties whom a resource helps only indirectly (e.g., a patient lives because his doctor gets the resource) should be given less weight than its direct effects, and (2) some differences in outcome between candidates may be irrelevant because achieving them is not the goal of the particular "sphere" which controls the resource (e.g., that one potential recipient in the health care sphere will write a novel if he receives a scarce drug should not count in favor of his getting it). Other differences in expected outcome between candidates may be covered by the Principle of Irrelevant Good, even if they are relevant to the sphere. For example, relative to the fact that each person stands to avoid death and live for ten years, that one person can get a somewhat better quality of life or an additional year of life should not determine who is helped, given that each wants what she can get. One explanation for this is that what both are capable of achieving (ten years) is the part of the outcome about which each reasonably cares most in the context, and each wants to be the one to survive. The extra good is frosting on the cake. The fact that someone might accept an additional risk of death (as in surgery) to achieve the "cake plus frosting" for himself does not mean that he should accept an additional risk of death so that another person who stands to get the greater good has a greater chance to live. For these reasons, Sobjectivity requires that we ignore the extra good in allocating, even though consequentialism and ex ante maximization of individual expected good would require us to do otherwise.

However, in life-and-death decisions, any *significant* difference between two people in expected life years may play a role in selecting whom to help. This result follows from the third form of Sobjectivity. Still, because the large additional benefit would now be concentrated in the *same person* who would already be benefited by having her life saved for at least the same period as the other candidate, the additional benefit should count for less in determining who gets the resource than if the additional benefit were distributed to a third person. This is

on account of fairness and the diminishing moral value of providing an additional benefit to someone who would already be greatly benefited. Large differences in expected quality of life among candidates for a resource should count in situations where improving quality of life is the point of the resource.

What if taking care of the neediest or most urgent conflicts with producing the best difference in outcome? Rather than always favoring the worst off, we might assign multiplicative factors in accord with need and urgency by which we multiply the expected outcome of the neediest and most urgent. These factors represent the greater moral significance of a given outcome going to the neediest (or most urgent), but the nonneediest could still get a resource if her expected differential outcome was very large.

We can summarize these views quantitatively in what I call an *outcome modification procedure for allocation*. If we first assign points for each candidate's differential expected outcome, we then assign multiplicative factors for need and urgency in accordance with their importance relative to each other and to outcome. We multiply the outcome points by these factors. The candidate with the highest points gets the resource.

Sometimes the conflict between helping different people can be reduced because it is possible to help everyone to some extent, even though not completely. Imagine a case where each stands to lose and gain the same thing and we can do one of the following: (1) certainly save five lives on one island; (2) certainly save one life on another island; or (3) reduce the chances of saving the five in such a way that all six now share the same reduced chance of being saved together. I argued above that a nonconsequentialist should prefer (1) to (2), but it is still possible for her to prefer (3) to both, at least (the suggestion is) so long as we reduce the chance of saving the majority by no more than the proportional weight (1/6) of the minority. There is a preference for (3) over (1), even though the expected utility of these two outcomes is the same, because all will now have a chance to share the same fate.

Finally, we should be aware that many real-life cases in which we can help everyone to some degree are even more complicated. A nonconsequentialist theory must deal with dividing resources among individuals who stand to lose and gain to different degrees, where the probability of satisfying the needs is different, and where the number of people who fall into different need/gain categories differs.

References

Anderson, E. (1993) *Ethics and Economics*, Cambridge, MA: Harvard University Press.

Bennett, J. (1981) "Morality and Consequences," in *The Tanner Lectures on Human Values*, vol. 2, ed. S.M. McMurrin, Salt Lake City: University of Utah Press, pp. 45–116.

Darwall, S. (1986) "Agent-Centered Restrictions from the Inside Out," *Philosophical Studies* 50 (3): 291–319.
Foot, P. (1978) "The Problem of Abortion and the Doctrine of Double Effect," in *Vices and Virtues*, Berkeley: University of California Press, pp. 19–32.
Foot, P. (1984) "Killing and Letting Die," in *Abortion: Moral and Legal Perspectives*, eds. J. Garfield and P. Hennessey, Amherst, MA: University of Massachusetts Press, pp. 178–85.
Fried, C. (1978) *Right and Wrong*, Cambridge, MA: Harvard University Press.
Kagan, S. (1989) *The Limits of Morality*, New York: Oxford University Press.
Kamm, F.M. (1985) "Equal Treatment and Equal Chances," *Philosophy & Public Affairs* 14 (2): 177–94.
Kamm, F.M. (1989) "Harming Some to Aid Others," *Philosophical Studies* 57 (3): 227–60.
Kamm, F.M. (1993) *Morality, Mortality*, vol. 1, New York: Oxford University Press.
Kamm, F.M. (1996) *Morality, Mortality*, vol. 2, New York: Oxford University Press.
Kamm, F.M. (2007) *Intricate Ethics*, New York: Oxford University Press.
McKerlie, D. (1997) "Priority and Time," *Canadian Journal of Philosophy* 27 (3): 287–309.
Nagel, T. (1986) *The View from Nowhere*, New York: Oxford University Press.
Quinn, W. (1993) "Actions, Intentions, and Consequences: The Doctrine of Double Effect," in *Morality and Action*, Cambridge: Cambridge University Press, pp. 175–93.
Rachels, J. (1975) "Active and Passive Euthanasia," *New England Journal of Medicine* 292: 78–80.
Scanlon, T. (2008) *Moral Dimensions*, Cambridge, MA: Harvard University Press.
Scheffler, S. (1982) *The Rejection of Consequentialism*, New York: Oxford University Press.
Sen, A. (1982) "Rights and Agency," *Philosophy & Public Affairs* 11 (1): 3–39.
Taurek, J. (1977) "Should the Numbers Count?" *Philosophy & Public Affairs* 6 (4): 293–316.
Thomson, J. (1985) "The Trolley Problem," *The Yale Law Journal* 94 (6): 1395–415.
Thomson, J. (1999) "Physician Assisted Suicide: Two Moral Arguments," *Ethics* 109 (3): 497–518.
Williams, B. (1981) "Utilitarianism and Moral Self-Indulgence," in *Moral Luck*, Berkeley, CA: University of California Press, pp. 40–53.

Further Reading

Foot, P. (1978) "Euthanasia," in *Vices and Virtues*, Berkeley: University of California Press, pp. 33–61.
Foot, P. (1983) "Utilitarianism and the Virtues," *Proceedings of the American Philosophical Association* 57 (2): 273–83.
Kamm, F.M. (1992) "Nonconsequentialism, the Person as an End-in-Itself, and the Significance of Status," *Philosophy & Public Affairs* 21 (4): 354–89.

Kamm, F.M. (1993) *Morality, Mortality*, vol. 1, New York: Oxford University Press.
Kamm, F.M. (2013) "The Trolley Problem," in *International Encyclopedia of Ethics*, ed. H. LaFollette, Oxford: Wiley-Blackwell.
Kant, I. (1964) *Groundwork of the Metaphysic of Morals*, trans. and ed. H. J. Paton, New York: Harper & Row.
Ross, W.D. (1930) *The Right and the Good*, Oxford: Oxford University Press.
Scanlon, T. (1982) "Contractualism and Utilitarianism," in *Utilitarianism and Beyond*, eds. A. Sen and B. Williams, Cambridge: Cambridge University Press, pp. 103–28.
Thomson, J. (2008) "Turning the Trolley," *Philosophy & Public Affairs* 36 (4): 359–74.

Chapter 13

Intuitionism

David McNaughton and Piers Rawling

Varieties of Moral Theory

What makes an action morally obligatory, the one that we are morally required to do? Different moral theories give different answers to this question. The simplest answer would be that just one consideration is relevant to the rightness of an action. *Act-consequentialism* (which we will refer to as "consequentialism," unless otherwise indicated) is a popular and influential theory that claims just this. On this view, the only morally relevant consideration is the effect an action will have on the amount of value in the world – you are morally required to bring about the greatest balance of good over bad.[1]

Consequentialists differ, however, as to what things are good. Some are *monists* about the good. Thus classical utilitarians, such as Bentham and Mill, asserted that there is only one intrinsic good: pleasure. *Pluralists* think there are a number of distinct goods. For example, it seems plausible that not only the amount of pleasure, but also the fairness of its distribution, is morally important. If there were two worlds containing equal amounts of pleasure, then the one with a fairer distribution would appear to be the better. Some other candidates for intrinsic goods are knowledge, beauty, achievement, and virtue.

Although consequentialists may be either monist or pluralist about the *good*, they can all be seen as monists about *moral obligation*, since they hold that only one consideration, the amount of value produced, is relevant to whether an action is morally required. They may allow what Mill called "secondary principles," such as "Do not tell lies" or "Keep your promises," but these principles derive their force from the primary principle: produce as much good as you can. And though

The Blackwell Guide to Ethical Theory, Second Edition. Edited by Hugh LaFollette and Ingmar Persson.

keeping your promises may generally produce more good than breaking them, and thus often be obligatory, there are going to be cases where this is not so, and then it would be obligatory to break your promise. *Deontologists* deny that moral obligation is a matter of value alone. A strong (or absolutist) deontological theory holds that, for at least some actions, the value of their performance is wholly irrelevant to the moral status of the act. For example, on many such views, killing the innocent is simply wrong, however much good it produces or bad it prevents. A rather weaker version admits that there may be occasions on which killing the innocent is permissible, or even required – to prevent other killings, say. But on such a moderate deontology, the threshold above which you should kill to prevent killing is above that which would be determined by value alone – for example, even if your killing one to save two would maximize value, the moderate deontologist might claim that you should not do it. Her moderation comes in when it comes to killing one to prevent catastrophe: here she thinks you should kill. One mark of deontology, in either form, is that it may often be obligatory to produce less good than one could. To take another example, deontologists hold that it is quite possible that you be morally required to keep a promise even though this would not maximize the good.

Intuitionism

Intuitionism is one among several different kinds of deontological theory. First, some offer a *monist* account of moral obligation, but intuitionism is *pluralist*. While most deontological theories accept that there are a number of moral principles or duties, monistic theories claim that these principles are grounded in, or derived from, some single principle, test, or principle of reason. Thus Kantianism starts from claims about the nature of practical reasoning, and from this derives a test (the Categorical Imperative) that any proposed principle (or maxim) of action must pass if it is to be morally acceptable. Our various duties are then generated by seeing which maxims pass or fail the test. Intuitionism, by contrast, claims that there are a number of distinct fundamental or underivative moral principles. They stand on their own feet, as it were: they are self-supporting.

Second, intuitionism holds that there are objective moral truths (it is a form of *moral realism*), and that these do not depend on our responses or attitudes. The intuitionist might agree, for example, with a certain kind of *response-dependent* account of morality, that an act is wrong if and only if it merits a certain kind of disapproval. But whereas the response-dependence theorist thinks that an act is wrong *because* it merits a certain response, the intuitionist sees the *direction of explanation* as reversed: the response is merited because the act is wrong.

Third, intuitionism is not a *constructivist* theory. Constructivists maintain that what makes a moral claim true is that we could arrive at it by following some valid procedure. Different versions suggest different procedures. Kantians, as we have seen, ask whether the principle on which we propose to act would pass the

Categorical Imperative test – roughly, could the principle be universally willed? Consider, for example, the principle "make lying promises (i.e., promises you have no intention of keeping) whenever it is to your advantage." This principle cannot be universally willed since if everyone were to make lying promises whenever they stood to gain by it, promising would cease to be possible – the practice would disappear. Contractualists require that we act only on principles that could not reasonably be rejected by people who share the aim of only acting in ways that they could justify to others. Intuitionists, by contrast, hold that the moral facts do not exist in virtue of any procedure we might have for uncovering them. Thus, again, there is a dispute over the direction of explanation. For a constructivist, moral principles hold because they can be arrived at by following the correct procedure. For an intuitionist, if a procedure is correct it will be so only because it reliably leads us to the moral facts, which obtain independently of the procedure.

Fourth, there is a division among realists between *naturalists* and *nonnaturalists.* Naturalists hold that the only things, facts, and properties in the world are, very roughly, ones that can be discovered by the empirical methods of science and that can play a role in scientific explanations. The intuitionists, following Moore's lead, held that moral claims are not empirical; they are not subject to scientific investigation. Moral claims are *normative*; and normative claims are distinct from empirical ones. The former, for example, tell us what we *should*, or *ought*, or have *reason* to do or think; the latter tell us how we *do actually* behave and think. The methods of science are well suited to establishing and explaining empirical claims, but not normative ones. For instance, science itself rests in part on normative claims of its own – scientists *should*, for instance, follow the scientific method. But this normative claim cannot itself be established by the scientific method; indeed, no amount of empirical investigation can, on its own, establish what we should do in any domain. Intuitionism holds that moral properties cannot be analyzed into, or reduced to, empirical properties. Historically, intuitionists disagreed, however, as to whether "good" and "obligatory" were themselves unanalyzable, or could be analyzed in terms of other normative properties.

To sum up: intuitionism is a form of nonnaturalist moral realism, which espouses pluralism about moral obligation. (Many intuitionists were also pluralists about the good, but that is not essential to their position.)

Intuitionism flourished most recently in England between the World Wars (although its roots go much further back). In the 1940s, it became unfashionable, and remained so until very recently. Few bothered to read its principal proponents – Prichard, Ross, Broad, Carritt, and Ewing – carefully or sympathetically, and consequently it was often caricatured. The very name "intuitionism" was a handicap, since it encouraged the popular misconception that the theory was committed to the existence of a mysterious faculty of moral intuition, unknown to science, by which we detect moral properties. In fact its proponents typically did not call themselves intuitionists, and the word "intuition" rarely occurs in their writings.[2] So where does the label come from? It was Sidgwick who coined the

term to denote the pluralist deontology that, he thought, characterizes our ordinary moral thinking, as opposed to the utilitarianism he advocated.

Rather confusingly (as Broad (1930: 206ff.) points out), Sidgwick also uses the term "intuitionism" to denote a method of knowing truths. Empirical truths are knowable by observation or experience, but others can be known a priori, without appeal to experience. Among the truths we can know a priori, some are deduced from other such truths, but if this process is not to go on forever, there must be some truths that are not deduced from others. These are self-evident: we can know them *intuitively*. Sidgwick argues that *all* ethical theories, and not just intuitionism, must have at least one self-evident principle, or intuition, at their base, since the foundational principles of a moral theory cannot, by their nature, be deduced from more basic propositions, nor can they be known by experiment or observation. What distinguishes intuitionism from utilitarianism (and consequentialism in general) is thus not the appeal to self-evident truths, which Sidgwick thinks it shares with all theories, but the fact that it is a pluralist deontology.

To exacerbate the confusion, the term "moral intuition" is nowadays used in a different way. Instead of denoting the grounding principle(s) of a moral theory, it is used to refer to our pretheoretical judgments about particular moral examples. For instance, most people have the moral intuition that a doctor might let one patient die in order to save five, but that he should not take organs from a living patient in order to save five who need transplants. Appeal to this kind of moral intuition also plays a role in all moral theories, and not just in intuitionism.

Objections to Intuitionism

One reason for dismissing intuitionism is the common conviction that it is fundamentally nonexplanatory and tells us nothing that we did not know already. While it is conceded that it gives a pretty accurate sketch of much of our everyday moral thought, "the theory, appraised as a contribution to philosophy, seems deliberately, almost perversely, to answer no questions, to throw no light on any problem. One might almost say that the doctrine actually consists in a protracted denial that there is anything of the slightest interest to be said" (Warnock 1967: 12–13). In short, it looks to its critics as if intuitionism should scarcely be dignified with the title of a theory at all.

So what should we expect from moral theory? Theories seek to give us a deeper understanding of their subject matter, and show us whether, and to what extent, various claims are justified. Moral theories have traditionally had at least four aims. First, they endeavor to systematize our moral thinking. Second, they address metaphysical questions: Are there moral properties? Are they like other properties? Third, they attempt to answer epistemological questions: Is there moral knowledge? How reliable are our moral beliefs? What methods are available for settling

moral disagreements? Finally, they aim to explain how moral knowledge is of practical, and not merely theoretical, significance: morality tells us how to live.

Intuitionism has often been thought to cast little illumination on any of these areas. First, intuitionism appears unsystematic: it portrays morality as an unconnected heap of duties with no underlying rationale. Second, intuitionists offer only a negative account of moral properties: they are not natural – they lie outside the compass of the empirical sciences. Far from increasing our understanding, this claim threatens to make matters more mysterious. Worse, it seems to undermine the possibility of moral knowledge. If we cannot study these properties by observation and experiment, how do we know about them? It looks as if the answer must be: "by intuition." But, as Warnock pithily remarked, this seems "not really an answer at all, but a confession of bewilderment got up to look like an answer" (Warnock 1967: 7). Lastly, even if we had such knowledge, intuitionism does not seem to explain how it could make a difference to how we live. How might awareness of these mysterious nonnatural properties play a distinctive role in determining how we should act?

We have seen, however, that even its critics concede that intuitionism does a good job of delineating the outlines of our ordinary moral thinking. This is certainly a strong point in its favor, since it is widely agreed that a moral theory should not deliver moral verdicts too far out of line with everyday moral judgments. If the four objections just sketched can be met, the theory will be in good shape. In fact, many of these criticisms rest on serious misunderstandings. To clear those up, we shall examine the views of W.D. Ross, who developed the best-known and most systematic version of the theory.

It should be conceded at the outset, however, that there is one respect in which intuitionism does not meet the aspirations of many moral philosophers. Their hope is that moral theory can *advance* moral knowledge by providing some definitive principle or procedure for answering some of the many puzzles and disagreements that beset our ordinary moral thought. Intuitionists, however, are skeptical about the pretensions of moral theory to resolve what our ordinary moral thinking leaves disputed or unclear. All but the most trivial moral decisions involve judgment, the comparison and weighing of competing considerations. And no abstract principle or procedure can obviate the need for judgment. Theory cannot make our weighing more precise, or our judgments more certain.

The Structure of Ross's Intuitionism

Intuitionism claims there are a number of distinct fundamental moral principles. That claim immediately raises two structural questions. First, what happens when there is a conflict between principles? Second, how many fundamental moral principles are there, and how do we decide what should be on the list?

Ross's Conception of "Prima Facie Duty"

Any pluralist moral theory that allows cases in which, whatever we do, we shall breach some principle, has to say something about such moral conflicts.

Some hold that a complete moral theory needs to have a higher-order rule or procedure that tells us what to do when lower-order rules conflict. Some theories (such as Kant's), for example, divide duties into two classes: those that admit of exceptions, and those that do not. In cases of conflict between duties of different types, the exceptionless principle wins out. This rule does not provide a full ordering of principles, but even this partial ordering comes at too high a price in Ross's view, for he denies that there are any exceptionless general moral principles. Whatever principle we are considering, there are always some circumstances in which it would be obligatory to breach it.

Furthermore, according to Ross, although there are some fundamental principles, they cannot be prioritized. Sometimes, for example, our duty to keep a promise should take precedence over our duty to help others, but on other occasions the opposite is true – it all depends on how serious the promise, how great the good to be achieved, and the context in which the conflict occurs. Ross expressed this thought – that no principle or duty systematically trumps another – by saying that the fundamental moral principles are *prima facie duties*. And, to emphasize a key point, there is no algorithm for determining what is obligatory – having worked out which prima facie duties apply in a given circumstance, there is the further question of which are the weightiest: what is morally required here? And responding to this question, on Ross's view, requires judgment: there is no mechanical decision procedure for generating the correct answer.

Ross's terminology is doubly misleading, as he acknowledged, since the features to which he was trying to draw attention are, strictly speaking, neither prima facie nor duties. To say that, prima facie, something has a certain characteristic suggests that it appears, at first sight, to have that characteristic, but that subsequent investigation might show that it does not. But this is not what Ross means. If an act is prima facie required it remains so even when, because of other facts about it, it turns out to be, overall, not obligatory.

Ross thought that the term "duty" was also misleading. Why? Because he holds that, strictly speaking, only the act that you ought to do, all things considered, is your duty. Talk of prima facie duties unfortunately suggests that "what we are speaking of is a certain kind of duty, whereas it is in fact not a duty, but something related in a special way to duty" (Ross 1930/2002: 20). These terminological difficulties can be avoided if we talk, instead, of fundamental morally relevant considerations or, even more simply, of fundamental moral reasons.

What exactly is the relation, then, between a fundamental moral reason and an action's being overall obligatory? Philip Stratton-Lake (1999) has suggested that we should think of the relationship as analogous to that between evidence and verdict. A moral reason would thus be analogous to a piece of evidence that must

be taken into account in reaching the final verdict on the obligatoriness of the act. This is true, as far as it goes, but the relation is closer than the terms "evidence" and "verdict" suggest. Something can count as evidence for a thing's having a certain property without its contributing to the thing having that property. That Simon says it is sunny is evidence that it is sunny, but his remark does not contribute to its being sunny (in the way that, say, a cloudless sky does). However, the moral reasons that count in favor of an obligatory act are not just reasons to *believe* that it is morally required; they *contribute* to its being so (see McNaughton and Rawling 2011).

In many cases, there will be moral considerations both for and against some action. Whether I ought to do the action depends on which set of reasons is the weightier. But even defeated considerations, ones that were on the losing side, retain their moral force and so can leave me with a residual obligation. Suppose I ought to break my promise to take my son to the circus in order to visit my sick mother. The fact that I have made a promise remains morally relevant, and can affect what I ought subsequently to do. The defeated reason is not removed or canceled. I should do something to make amends to my son for the broken promise (Ross 1930: 28).

Determining Which Moral Considerations Are Fundamental

Our ordinary moral thinking appears messy and unstructured. People appeal to a plethora of principles in trying to justify their views on some specific moral issue. They may appeal to substantial moral principles such as "Do not tell lies," or to formal procedural principles such as "Do as you would be done by," or to some combination of the substantial and the formal. All moral theories attempt to delve beneath the surface and uncover the underlying structure of our moral thought. In so doing, they hope not only to explain our moral judgments but also to reveal whether, and to what degree, they are justified. One criterion for a good theory is simplicity. So, everything else equal, the fewer principles that we need to invoke the better. But parsimony is not the only virtue, so a theory that has too few principles to explain our moral thought is inadequate. Ross's intuitionism shares this methodology with the other main moral theories; the disagreement concerns both the number of principles we need and their nature.

Ross famously offers the following division of prima facie duties (1930/2002: 21). It is not an arbitrary list but is intended as a first shot at a list of moral considerations that are both fundamental and distinct from each other. There are seven:

(1) keeping promises or commitments (fidelity)
(2) making amends for my wrongdoing (reparation)
(3) benefiting those who have benefited me (gratitude)
(4) distributing benefits and burdens according to merit or desert (justice)

(5) bettering the condition of others (beneficence)
(6) bettering my own condition (self-improvement)
(7) not harming others (nonmaleficence).

To offer such a putative list of fundamental morally relevant considerations invites, of course, the following question: what is it to be a derivative consideration as opposed to a fundamental one? Take the duty to obey the laws of one's country. Ross suggests that it "arises from" three basic duties: gratitude, fidelity and beneficence. Normally, we owe a debt of gratitude to our country for benefits received; we have made an implicit promise to obey its laws by living in it; and we should be law-abiding because things go better for society if we are. Similarly, there are two fundamental principles that underlie the duty not to lie: nonmaleficence and fidelity. Lying normally inflicts an injury on the person lied to, and undermines an implicit undertaking, underlying day-to-day communication, to tell the truth (Ross 1930/2002: 54–5).

These examples suggest the following picture: It is the mark of a derivative prima facie duty, such as the duty not to lie, that it can be explained in terms of some more fundamental feature(s), from which it can be said to "arise." Such explanations cannot go on forever. Eventually we will reach the bedrock of underivative duties; here we cannot appeal to anything more basic to explain why the features they mention are morally significant. What is fundamental is morally significant in its own right, Ross thinks, and thus always carries moral weight.

Ross makes it clear, in discussing his examples of lying and lawbreaking, that there can be special circumstances in which one or more of the fundamental considerations which normally count against acting in these ways do not obtain. In such cases, the force or bindingness of the derivative duty may be weakened. Ross holds, for example, that the implicit undertaking not to lie cannot hold where I am a complete stranger in another society and have had no chance to reach agreements of any sort with its members. Since he holds that a large part of our duty not to lie stems from the supposed implicit promise, its absence greatly weakens our duty not to lie.

Although Ross does not discuss this point, it seems perfectly possible that there might be cases where *none* of the considerations which normally tell against lying or lawbreaking apply. If we are playing a game like Cheat, then we have all agreed to suspend the normal conventions about truth-telling; far from people being harmed, the lying adds to the fun. A government may be so oppressive and unjust that neither gratitude nor beneficence dictates that its citizens should obey. And if, in addition, it refuses to let dissidents emigrate, then any argument from tacit consent to the government also lapses. This might suggest that, while being in breach of a fundamental prima facie duty *necessarily* counts against an act, being in breach of a derivative duty only counts against it contingently. Thus lying, for example, is sometimes not even prima facie wrong.

Are all derivative duties contingent in this way? Maybe not. Take the duty to pay one's debts. It seems that you can be indebted only in one of two ways: either

you have borrowed something on the understanding you will pay it back, or someone has done you a good turn. If that is right, then being indebted will always be morally relevant, but, because the duty to pay debts always falls under the basic duties of either fidelity or gratitude, it does not appear on the list of fundamental duties.

To return to this list, having sorted out to some extent what it is to be "fundamental" as opposed to "derivative," we can now ask the following pair of questions: First, is everything on the list fundamental, or can the list be shortened? And, second, are there any other items that need to be added – are there "duties" that do not arise from items already on the list? With regard to the latter question, Ross does not claim completeness for the list (1930/2002: 23). And regarding the former, Ross himself thinks that he can shorten his original list. Ross is a pluralist about the good, and he holds that self-improvement, justice, and beneficence reflect its components – so we could, then, replace this trio with a single prima facie duty to "produce as much good as possible" (1930/2002: 27) (although we would, of course, lose information by doing so). We turn now to the question of why *all* of Ross's duties cannot be subsumed under this single duty to produce good – or, to put it another way, why is Ross not a consequentialist?

The Rejection of Consequentialism

There are four other duties on Ross's list, so the issue is: why cannot these be subsumed under the duty to produce good? The first three – fidelity, reparation, and gratitude – are what are sometimes called "duties of special relationship." In each case the duty rests on some previous act either of my own or of others. It is because something morally significant has already occurred in the relationship between us that the other person has a claim on me. The other person is not just an instance of someone whom I could benefit. Rather, she has a claim on me, in virtue of our relationship, to a benefit, and often to a very specific benefit. Others who do not stand in these relationships do not have these claims. As Ross says:

> The essential defect of the 'ideal utilitarian' theory [i.e., consequentialism] is that it ignores, or at least does not do full justice to, the highly personal character of duty. If the only duty is to produce the maximum of good, the question who is to have the good – whether it is myself, or my benefactor, or a person to whom I have made a promise to confer that good on him, or a mere fellow man to whom I stand in no such special relation – should make no difference to my having a duty to produce that good. But we are all in fact sure that it makes a vast difference.
>
> (1930/2002: 22)

Ross holds not only that, given a choice, I should benefit my benefactor, or the person to whom I have made a promise, rather than a stranger, but that I should

do so even if I could benefit the stranger slightly more at no extra cost to myself or others.

> Suppose . . . the fulfilment of a promise to *A* would produce 1,000 units of good for him, but that by doing some other act I could produce 1,001 units of good for *B*, to whom I have made no promise, the other consequences of the two acts being of equal value.
>
> (1930/2002: 34–5)

In such a case we would, Ross claims, believe we should keep our promise. If we are right in this belief, he holds, consequentialism must be false, since we ought to produce the state that has less value.

But is Ross's reasoning correct here? We think not. He is assuming that, when we compare the value of two states of affairs, we need only consider the size of the benefits to all individuals. So the state that contains more benefits will contain more value. But consequentialism need not assume this, as Ross should know. For, on his own account of justice, a state of affairs in which benefits are fairly distributed is more valuable than one where the same amount of benefit is unfairly distributed. So, a consequentialist could maintain that acts of promise-keeping, gratitude, and reparation are themselves valuable. In that case the total value of keeping the promise to *A* will be the sum of the value that comes from *A*'s receiving those benefits, plus the value of promise-keeping itself. In which case, keeping the promise to *A* might well produce more value than giving a slightly larger benefit to *B*.

In reply, the Rossian could concede this point and still maintain that act-consequentialism is mistaken about both the strength and the nature of duties of special relationship. It underestimates the weight we think these obligations have. Suppose I have promised to pay back money I have borrowed, but decide instead to donate that sum to famine relief. Many people would think I acted wrongly, although it is plausible to argue that I would bring about more good by feeding the starving, even when we take into account the value of promise-keeping. This comes out most dramatically if we envisage a case where I encourage other people to keep their promises, but at the expense of not keeping my own. Even if I am fairly persuasive, so that a few more promises are kept as a result of my endeavors than would be if I focused on keeping my own promises, it seems implausible to claim, as act-consequentialism would, that to keep my own promises would be wrong.

It might be argued that consequentialism underestimates the strength of duties of special relationship, because it misunderstands their nature. As well as any general duty I may have to encourage acts of promise-keeping, I also have a specific duty *to my promisee* to keep my promise. It is precisely the personal nature of such duties that consequentialism cannot allow to be intrinsically morally significant. The consequentialist claims that I should adopt a policy of keeping my promises, but only because such acts tend to make the world a better place.

However, it is not just that the world is better for people honoring their commitments; the people to whom we are committed have special claims on us.

What about the duty of nonmaleficence – the duty not to harm? Can consequentialism accommodate it? It depends on how we understand this duty, and here Ross is somewhat unclear. He claims that it is wrong to inflict harm on someone in order to produce a similar, or slightly larger, benefit for someone else. But this might be because bringing about a harm of a certain sort produces more disvalue than failing to give a benefit of a similar sort. Generally speaking, taking away something someone already has seems far worse than failing to give him that thing when he lacks it. If that is so, then consequentialism can accommodate the thought by simply recognizing the greater disvalue in depriving someone of an existing benefit.

But there may be more to the duty not to harm than this. We tend to think it wrong to harm directly some (innocent) person, even to prevent someone else harming another innocent person. This might suggest, though the issue is disputed, that there is (what has come to be known as) a *constraint* against harming: the fact that the act would involve *my* directly harming another person is a sufficient reason for me not to do it, even if the disvalue that results if I refrain from harming would be the same, or somewhat greater. There are things we owe it to others not to do to them, even to prevent other people doing similar awful things.

Ross is opposed to consequentialism, then, insofar as he believes that there are duties of reparation, gratitude, fidelity, and nonmaleficence, which sometimes require us not to bring about as much good as we can. But in another respect Ross agrees with consequentialism, for he holds that, all else equal, we should produce as much good as we can. And that leaves him open to some of the criticisms that have been leveled against consequentialism. Most obviously, his view makes morality very demanding. We typically think that there are saintly or heroic actions (supererogatory acts) that go well beyond the call of duty and for which people receive especial praise. But on his account, since we should do as much good as we can (provided other duties do not forbid us) heroes and saints are only doing their duty. Commonsense morality seems to allow us *options* (as they are sometimes called) – permission not to do the most good if it would require too great a sacrifice. Though Ross thinks that consequentialism tells only part of the story, he does think that it gets its part of the story right, and that brings his theory into uncharacteristic conflict with normal reflective ethical thought (see Wiggins 1998; Dancy 1998, and Darwall 1998).

Ross's theory, like consequentialism, overmoralizes our choices. As we have seen, there is always something we are required to do, since every action affects value, and we have a prima facie duty to maximize value. Further, the only consideration that can outweigh a prima facie duty is another prima facie duty. But that is doubtful. One can on occasion have good reasons not to fulfill a prima facie duty which are not themselves reasons of duty. I promised to mark this student's essay by tomorrow. I am very tired and nothing disastrous will happen if I am a day late. These are good reasons to go to bed rather than mark the essay.

But should we think of them as constituting a duty to go to bed – one that is weightier than any duty we have to the student? That seems excessively moralistic. Given the effort involved, it might seem that we should again think in terms of options: I am not required to keep my promise if it is proportionately too costly.

These are genuine worries about Ross's system, but easily accommodated by a friendly amendment. If we distinguish between moral reasons and other types, we can allow the possibility of a nonmoral reason being morally relevant. The cost to me may be a reason not to grade the essay, but it is not a moral reason. What I should do is what I have most reason to do. It is where I have most reason to do this act, and (roughly) the preponderance of reasons in favor of my doing it are moral ones, that I can be said to be *morally* required to do it.

The Rejection of *Rule*-Consequentialism

So far our discussion of consequentialism has focused upon *act*-consequentialism, according to which we have one moral duty: to maximize the good. But there are other versions, one of which, rule-consequentialism, might appear to be rather close to intuitionism. According to this form of consequentialism:

> An act is wrong if and only if it is forbidden by the code of rules whose internalization by the overwhelming majority of everyone everywhere in each new generation has maximum expected value in terms of well-being (with some priority for the worst off). The calculation of a code's expected value includes all costs of getting the code internalized. If in terms of expected value two or more codes are better than the rest but equal to one another, the one closest to conventional morality determines what acts are wrong.
>
> (Hooker 2000: 32)

It may be that the rules that make up the proposed code are fairly similar in content to Ross's prima facie duties. And assuming that this is the case, rule-consequentialism might seem to have an advantage over intuitionism: its rules have an underlying, and unifying, justification that intuitionism lacks – namely, that the internalization of the rules would maximize expected value (see, e.g., Hooker 2000: 107). But this purported advantage might be outweighed by the various problems facing rule-consequentialism. Here we shall focus on a difficulty that stems from the fact that rule-consequentialism has to take into account the costs of inculcating the rules.

According to intuitionism, for example, we have fundamental duties of fidelity and gratitude toward our friends. The rule-consequentialist has to ask whether rules concerning loyalty and gratitude maximize expected utility – perhaps the world would be better if we were less partial toward our friends and family. Suppose, indeed, that a world in which we treat everyone impartially would be

better than one in which we display partiality, leaving considerations of cost aside. We do not endorse this view; but the point is that the rule-consequentialist has to include such costs in their calculations – and Hooker concludes that they "outweigh the benefits" of an impartial world: "Imagine what psychological and financial resources would have to be devoted to getting the overwhelming majority of children in each new generation to internalize an overriding equal concern for all others! The costs would outweigh the benefits" (Hooker 2000: 141). These cost considerations are completely alien to the intuitionist perspective. Fidelity, for instance, is basic. It is not something that has to be justified by appeal to something more fundamental – and certainly not on the grounds that it is too expensive to stamp out.

And rule-consequentialism's requirement that inculcation costs be considered leads to further implausible consequences. Imagine, for example, a world in which people take great pleasure in torturing cats, and in which it would be so expensive to get the population to internalize a rule prohibiting this practice that the costs would outweigh the benefits. Thus, according to rule-consequentialism, cat torturing would be morally permissible in such a world. This, of course, the intuitionist adamantly denies. A key difference between intuitionism and rule-consequentialism, then, is that the rules of the latter are contingent upon various things, such as internalization costs, whereas Ross's prima facie duties are not: cat torturing is just plain wrong, according to intuitionism, no matter how hard it is to resist the temptation.

Intuitionist Epistemology

Methodology

In developing his moral theory, Ross employs what came later to be called the method of reflective equilibrium. We look both at our intuitions about particular cases, and at plausible general principles. Where they conflict about what we ought to do, we make adjustments to each until they are in accord with each other. Ross's particular slant on this procedure is that our reflective judgments about particular cases are to have the final say. We do, of course, form snap judgments about particular cases, and we may revise these not only in the light of judgments about other cases, but also in the light of plausible principles. If, however, after careful reflection, our judgment about a particular case remains at odds with the principles, it is the latter that must give way. If, for example, consequentialism tells us that

> we should give up our view that there is a special obligatoriness attaching to the keeping of promises because it is self-evident that the only duty is to produce as much good as possible, we have to ask ourselves whether we really, when we reflect, *are*

> convinced that this is self-evident, and whether we really *can* get rid of our view that promise-keeping has a bindingness independent of productiveness of maximum good. . . . To ask us to give up at the bidding of a theory our actual apprehension of what is right and what is wrong seems like asking people to repudiate their actual experience of beauty, at the bidding of a theory which says 'only that which satisfies such and such conditions can be beautiful'.
>
> (Ross 1930/2002: 40)

So, what is the nature of this "actual apprehension," and in what cases can we have moral knowledge?

Certainty and Probable Opinion

Ross draws a sharp distinction between "our apprehension of the *prima facie* rightness of certain types" of action (1930/2002: 29) and our judgment about the overall obligatoriness or wrongness of particular acts. Claims about prima facie obligatoriness (or wrongness) are self-evident. That an act is, for example, prima facie required in virtue of being the keeping of a promise is something that we can know a priori, by reflection – and it is something of which we can be certain. By contrast, about what we should do – what is overall required – in some particular case we can only form what Ross calls a "probable opinion." The main reason for this is that there are nearly always morally relevant considerations for and against any act (and even when we do not think there are, we cannot be sure). To decide what to do, we have to balance those considerations, and that calls for the exercise of judgment – as noted in the subsection "Ross's Conception of 'Prima Facie Duty'," there is no algorithm for determining which of our prima facie duties are the weightiest in any given case.

Self-Evidence

The fundamental prima facie principles can, as noted, be known with certainty. But how? Initially, we may accept that we should, say, keep our promises on the basis of authority. But, as we mature and reflect upon our judgments in various cases concerning promise-keeping, we can come to see for ourselves that we have a prima facie duty to keep them (see Ross 1939: 170–3). That we have such a duty is a self-evident truth – it requires no proof. Rather, one can know a self-evident truth on the basis of simply understanding it (see Audi 1996: 114).

What is self-evident need not be *obvious*. Such propositions are evident to those with sufficient mental abilities and experience who have reflected properly about them. Ross's analogy here is with our knowledge of mathematical axioms and forms of inference. The fundamental principles of mathematics, logic, and ethics are not *analytic*; that is, they are not true in virtue of the meanings of the terms

employed in them. They are, thinks Ross, synthetic propositions that can be known a priori. Whether and how there can be synthetic a priori knowledge is a contentious issue but – and this is the important point in defending intuitionism against its detractors – Ross is not claiming that moral principles are known by some special and mysterious faculty, as his detractors suppose (see Mackie 1977, and also Audi 1996).

There are, of course, many differences among mathematics, logic, and ethics. But they have this in common: their fundamental propositions are not susceptible to empirical justification – a feature they share with philosophy, and, indeed, empirical science itself. The claim that the hypothetico-deductive method leads to knowledge, for example, is not itself susceptible to justification on empirical grounds. The idea, then, is that these fundamental propositions can neither be verified empirically, nor derived from anything more basic. If they are known, it must be through reflection on their contents: they are self-evident. Ross is here placing himself squarely in a mainstream philosophical tradition that holds there are substantial claims, including ethical ones, whose truth we can know by direct rational insight.

Is this line correct? Robert Audi suggests that Ross sometimes expresses himself in ways that make his claim sound stronger, and thus less plausible, than it need be. It is not, for example, necessary in order to apprehend the truth of a proposition that is self-evident, that one also recognize that it is self-evident (Audi 1996: 106). Nor should we follow Ross in claiming that we can be certain of the general principles of duty if by that he meant that we could not be mistaken about them. Moore was undoubtedly right when he said that by calling some propositions intuitions he (Moore) means

> *merely* to assert that they are incapable of proof. . . . [I do not] imply . . . that any proposition whatever is true, *because* we cognise it in a particular way. . . . I hold, on the contrary, that in every way in which it is possible to cognise a true proposition it is also possible to cognise a false one.
>
> (Quoted in Audi 1996: 108)

Further reflection on what seems self-evident can lead to a change of mind.

As Audi further points out (1996: 117), Ross (and Moore) also make a stronger claim than they need in claiming not only that self-evident principles need no proof, but also that they cannot be proved. Though we can know them without evidence, this does not mean that there could be no further evidence for them. True basic propositions are self-evident, but self-evident truths need not be basic – indeed, perhaps self-evidence is relative to the knower. When it comes to the items on Ross's list, however, we are more skeptical than Audi that any further theoretical underpinning can provide them with independent support. Some have worried that intuitionism lacks the unity desirable in a theory because the basic duties are not connected to each other. Audi suggests that we might see them all as, for example, expressions of respect for persons. But we doubt that this proposal

provides a unity that is more than merely verbal, since what it is to respect persons, one might think, is spelt out in terms of Ross's disparate duties.

What is the Role of the Prima Facie Duties?

The prima facie duties clearly play a classificatory role – acts can be classified as harmful or beneficent, as the keeping of a promise or the repayment of a debt of gratitude. But do they play a larger role than this? In particular, do they help guide us in determining our moral obligations? Ross thinks that, in general, they do not. While there may be special circumstances in which we have to consult the list of duties in order to see what we have moral reason to do, on most occasions we can see directly the moral pros and cons of any proposed course of action. (And, as already mentioned, judgment is then required to assess what is required of us overall.) It is, then, on Ross's view, true that we have the prima facie duties we do, but their chief role may be merely classificatory. We shall return to this issue when we discuss particularism in the final section "The Place of Moral Principles: Generalism or Particularism?"

Normativity, Motivation, and Practical Reasons

Moral beliefs have practical import: they should and do make a difference to how we live. Some think that intuitionism has special difficulty explaining this. If moral facts are facts like any other, why might not someone notice them but pay them no heed?

There are two distinct issues here. The first, normative, question concerns why we *should* be guided by moral considerations; the second, motivational, question concerns how moral beliefs motivate. Normative realists hold that, in addition to nonnormative facts – such as the fact that it is cold – there are normative facts, such as the fact that it is cold is a reason for you to wear your coat. Normative facts are facts about what we ought to, or have reason to, believe or do. Theoretical deliberation concerns what we have reason to believe; whereas practical deliberation concerns what we have reason (including moral reason) to do. On Ross's view, facts about what we have a duty to do (whether prima facie or all things considered) are normative facts: facts about what we have moral reason to do. So the old question "Why should we be moral?" answers itself. It is pointless to ask what reason there is to do what we have reason to do.

What about moral motivation? *Externalists* about moral motivation see it as possible that someone might believe she has a moral reason to act in a certain way, and yet not be motivated in the least so to act. *Internalists* deny this. Intuitionism is consistent with either view. Suppose I believe that Eve has a headache and an aspirin will relieve it. And suppose I also believe that the fact just cited is a moral reason for

me to give Eve an aspirin. On the externalist view, I will be motivated to give Eve an aspirin just in case I have a desire (which I might lack) to do what I take myself to have moral reason to do. On the internalist account, by contrast, I could not believe that I have moral reason to give Eve an aspirin unless I were also motivated to do so. Intuitionism does hold that there are two (distinct) beliefs here: my belief that Eve has a headache and an aspirin will relieve it, and my belief that this is a moral reason to give her one; and the view also holds that these beliefs play a crucial role in motivation. But the intuitionist can remain silent on the further details of how my motivation to give Eve an aspirin arises (if it does).

This does not mean, however, that no such details can be given, or that intuitionism does not give an outline of the overall story. If I am functioning well, I have the beliefs in question because it is true both that Eve has a headache and that this is a reason for me to give her an aspirin; and these truths motivate me to give Eve an aspirin. Internalism simply limits the possibilities for motivational malfunction: I cannot fail to be motivated to give Eve an aspirin if I believe I have reason to do so.

There is, however, a further issue here, concerning the direction of explanation. For the intuitionist, reasons, and the fact that they are reasons, explain agents' beliefs that they have reasons to do things; and these beliefs play a crucial role in agents' motivations. When an agent sees the world aright, she is motivated to act in some way because she has reason to. But on certain Kantian accounts, by contrast, an agent, if rational, has a reason to do some act only when, and because, she is motivated to do it.

In the section "Intuitionism," we briefly discussed constructivism, and pointed out that Kantian constructivists see moral principles as holding because (roughly) they pass the Categorical Imperative test. But some Kantians also see a role for motivation in determining moral principles. For example, according to Darwall (2006, in Copp 2006: 299): "In [the Kantians'] view . . . something's standing as a normative reason ultimately depends on its being motivating (treated as a reason) in fully rational deliberation, where the latter is determined by internal, formal features of the deliberative process, not by its responsiveness to independently establishable normative reasons." So, according to Darwall's Kantian, that, say, an utterance would be a lie is a reason not to utter it because any fully rational agent is motivated not to lie (where to be fully rational, in part at least, is to act only on principles that pass the Categorical Imperative test).

The intuitionist, on the other hand, claims that there *are* "independently establishable normative reasons." Suppose that if A were to utter S he would be lying; and suppose further that this is a reason for A not to utter S. If A is virtuous (or fully rational, in Kantian terms) and grasps these facts, he will be motivated not to utter S. But he will be so motivated because he has reason not to utter S and sees this – on the intuitionist account the fact that an idealized agent is motivated not to utter S does nothing to make it the case that there is a reason not to utter it.

The intuitionist takes there to be normative facts – about what our prima facie duties are, about whether an act falls under any of them (hence about what you

have moral reason to do), and about what we are morally required to do. We arrive at knowledge of these facts through reflection upon, and wise judgment about, cases, both actual and possible. From the Kantian perspective, however, these judgments comprise unsubstantiated moral assumptions. The application of the Categorical Imperative, on the other hand, requires no substantive – that is, moral or evaluational – input: it is a formal rather than a substantive procedure. And while moral judgment at some points is required on the Kantian approach (for example, it is a Kantian requirement that we help strangers, and judgment is required to determine how and when to do this), applying the Categorical Imperative does not require the making of moral judgments. On the intuitionist approach, by contrast, moral judgment is required at all points. There is no procedure, let alone a formal procedure, for determining what our prima facie duties are, which of them an act falls under, or how to weigh their relative importance in cases of conflict.

The intuitionist, in addition to seeing the Kantian procedure as putting the cart before the horse when it comes to the direction of moral explanation (see the section "Intuitionism") and of motivational explanation, also sees the Kantian procedure as often questionable in its application and incorrect in its verdicts. And even when the verdict is correct, the procedure is in danger of failing to furnish the correct reason for action. Suppose that the principle "harm the innocent whenever it is to your advantage" fails the Categorical Imperative test. Surely this failure is not itself the reason not to harm the innocent. Nevertheless, many critics see intuitionism as insufficiently "principled," and, from this perspective, that at least Kantians are proposing a decision procedure. And even if Kantianism fails in its current form, perhaps, with sufficient ingenuity, it can be saved (see, e.g., Parfit 2011, pts 2 and 3). (Of course, the intuitionist claims that such principles must conform to our antecedent reflective judgments, so that they are at best redundant.)

From the Kantian side, then, intuitionism is attacked for its overreliance on judgment, and lack of formal procedures. However, the intuitionist does subscribe to some principles – the prima facie duties. We conclude with an attack from the other direction: some see intuitionism as *too* principled.

The Place of Moral Principles: Generalism or Particularism?

What is the role of moral principles in ethics? There is a spectrum of views. At one end there are those who hold that the job of moral philosophy is to refine our principles so that they can be used with precision to deliver clear verdicts. This is to understand morality by analogy with law (as the common phrase "the moral law" implies). Written laws try to define as precisely as possible exactly what actions constitute a legal offence. And where statutes fail in this regard, judicial decisions make laws more precise and set a precedent that guides later judgments. At the

other end, extreme particularists (as they are known – the coinage is Hare's, 1963: 18) have denied the very existence of useful moral principles. Each particular case is different, and appeal to principles, it is feared, oversimplifies and distorts the nature of moral thinking, downplaying the role of imagination and judgment in the morally sensitive person's assessment of what should be done in *this* case. This is to understand morality by analogy with aesthetic judgment. Attempts to lay down rules of aesthetic appreciation appear ridiculous, and each work of art has to be judged on its own individual merits.

Where should we place Ross on this spectrum? While he is not a particularist since principles play a role, that role is significantly limited, in a number of ways. First, as we have seen, Ross holds that appeal to principles cannot settle what to do in conflict cases, and for Ross conflict is the norm, since there is nearly always something to be said on both sides. Ross offers very little general guidance for resolving conflicts; he offers a few remarks about the comparative stringency of the prima facie duties – that fidelity is usually a more weighty duty than beneficence, for example – but that is all. For the rest, Ross says, citing Aristotle, the decision rests with perception, that is, sensitivity to the details of the particular case: "This sense of our particular duty in particular circumstances, preceded and informed by the fullest reflection we can bestow on the act in all its bearings, is highly fallible, but it is the only guide we have to our duty" (1930/2002: 42).

Even deciding whether an act falls under one of Ross's prima facie duties often requires judgment. Take nonmaleficence. Philosophers have spilled much ink trying to provide a watertight account of when someone has violated a duty not to harm. Ross seems to have little interest in attempting to make the harm principle more precise. That may be because the notion of harm itself eludes codification. (We return to this point shortly.)

Furthermore, Ross's methodology, as we have seen, is "bottom-up" rather than "top-down." A top-down theory, such as Kant's, holds that we start by appeal to plausible abstract principles which we then use to determine what is morally relevant and which types of act are required, permissible, or forbidden. For Ross, justification goes in the reverse direction. We frame our principles in the light of what we judge about particular cases. Indeed, Ross holds that basic principles play no role in helping us understand the moral character of an act. He asks: "Once the general principles have been reached, are particular acts recognized as right by deduction from general principles, or by direct reflection on the acts as particular acts having a certain character?" (1939: 171). He claims that the latter is virtually always the case. But if I can tell straight off, without appeal to principles, that, say, its cruelty counts against this action, the principles do no epistemic work.

Particularism began as a challenge to the claim that there are moral principles of the kind Ross proposes. Certainly, some of Ross's own principles seem vulnerable to counterexample. Does a promise to do a wicked deed give me *some* reason, however weak, to do it? If my benefactor helped me by perpetrating some horrendous crime, do I have any reason to be grateful? In just punishment, does not the fact that it will harm the guilty person count as a reason in favor of inflicting

the penalty, rather than against? Those who think that there must be general moral principles will respond by insisting that the principle needs more careful formulation to make it immune to counterexample: there is moral reason not to harm the *innocent*; promises are only binding under certain conditions; and so on. There is a danger of a dialectical standoff here. The particularist will try to provide ingenious counterexamples to the qualified principles; the principlist will seek to add yet more riders and qualifications to deal with those counterexamples. This process can go on indefinitely without a decisive victory by either side.

Attention has increasingly turned, however, to the question of whether there *must* be principles, even if it is difficult to formulate them precisely. We have seen that Ross denies that principles are epistemically required, but he may think of them as an ineliminable feature of the moral landscape nonetheless. According to the particularist, whether or not you think there must be principles depends on whether you have a holistic or an atomistic conception of reasons. The holist contends that a consideration that counts in favor of an action in one circumstance can be irrelevant, or even count against it in another. Every act has a unique context, and, says the holist, the reason-giving force of any given consideration is dependent upon other considerations present in the case, where this dependence defies codification. What we might call, to employ a chemical analogy, the valence of any consideration can switch from case to case. So, for example, that an act will give pleasure has positive valence when the pleasure is innocent, but negative valence when the pleasure is sadistic. It is generally accepted that many reasons change valence in this way, but the atomist holds that, at the level of fundamental considerations, valence must be unchanging. And this may be Ross's view. In subsection "Determining which moral considerations are fundamental" we attributed to Ross the view that fundamental considerations are morally significant in their own right and thus always carry moral weight. That an act would violate fidelity, for instance, is a bedrock proposition in that there is no more fundamental duty from which that of fidelity arises – and this entails, on the view we are attributing to Ross, that violating fidelity has invariantly negative valence (that an act would violate fidelity always counts against it). But does this entailment hold?

We think not. That we have arrived at bedrock when we see that some act would violate fidelity does not entail that fidelity violation *always* counts against an act. The fact that we cannot, as it were, dig any deeper does not entail that the adjacent terrain makes no difference. Perhaps an analogy from epistemology might help. You declare that this shirt is red on the basis that it appears so to you. And there is nothing more fundamental to be said. But this does not entail that seeming red always indicates redness. Whilst nothing more fundamental can be said than that the shirt appears red, there are plenty of "surrounding features" that are relevant – such as the fact that you are not wearing color-distorting spectacles. But if asked the reason why you think the shirt is red, you would simply say "it looks red"; you would not attempt to run through all the possible relevant side conditions, such as, "I am not wearing color-distorting glasses; I am not shining a red light on the shirt; and so on and so forth." Ross, on this portrayal, is tacitly

assuming that because a consideration is fundamental it is therefore morally significant in its own right, in the sense that its reason-giving force is completely independent of any other facts of the case in question. But this assumption is subject to challenge.

We agree with Dancy (2004: 119), then, that we have yet to see a good argument in favor of the claim that there *must* be invariant reasons. But that leaves open the question of whether there actually are any. It is useful at this point to distinguish between two different kinds of principle, depending on whether they are framed in evaluative or nonevaluative terms. The principle "Do not tell untruths" is framed nonevaluatively: it is possible to determine whether some statement is untrue without reference to any value. But the duty of nonmaleficence employs the notion of harm, which is ineliminably evaluative or normative: to harm someone is not merely to hurt someone; it is to make them *worse* off. The original thrust of holism was to deny that there are *nonnormative* features that have invariant valence. This seems plausible: it is hard to think of a nonevaluative feature that always counts for (or against) an action. Take pain, for example. At first thought, it might seem that the fact that an act would cause someone pain always counts against her doing it. But what about masochists? So then we might try: the fact that an act would cause someone a sensation she dislikes always counts against her doing it. But what is to count as a sensation? Suppose his success galls her; is this a reason for her to thwart him? Well, it certainly is not a moral reason to do so. But what about inflicting pain on someone else? Is not the fact that your act would do this always a moral reason against your doing it? Well, what about the case of justified punishment? The fact that the felon does not enjoy imprisonment is part of the reason *for* imprisoning him.

But even if nonnormative features lack invariant valence, the contextual variance might be limited, so that any exceptions could be incorporated into moral principles (painful sensations count against unless . . .). But the holism particularism espouses is more radical than that – it claims that for any finite principle we come up with, we cannot know in advance whether there might not be an unforeseen context in which it fails to deliver the correct verdict.

While we are sympathetic to this claim that there are no *nonnormative* features that have invariant valence, there may nevertheless be *normative* features that do. So, for example, that an act would be just is, it would seem, always a reason for doing it. It is noteworthy that Ross casts all his moral principles in normative terms: fidelity, gratitude, reparation, not-harming, justice, beneficence, self-improvement. That is why moral sensitivity is required to apply them correctly. But it also seems plausible that at least some of them do pick out features with invariant moral valence. Although we have not the space here to argue this in detail here, the basic idea is that Ross is attempting simply to describe and categorize the different kinds of moral reasons – it is, as it were, built into the prima facie duties that they present us with reasons. To say that an act would be beneficent, for instance, is simply to say that you have a certain kind of reason to perform it – namely, it would make the world better by increasing someone's well-being.

It may be the case, then, that, as far as moral reasoning goes, morality might have been completely unprincipled – principles are not epistemically necessary. Indeed, as we have noted, Ross himself insists that we are capable of picking out what is morally relevant in the particular case without appeal to general principles, thus he has no need to claim that such principles *must* be available for this purpose. But if Ross is correct, it is nonetheless true that we always have reason, say, to repay debts of gratitude.

Suppose, however, it turns out that Ross is not correct. Will not this weaken the case for intuitionism, since the defense against the charge of being unsystematic rested on the claim that Ross's theory, like all other theories, sought to uncover the most basic principles from which all others were derived? Drop that claim and you abandon the defense. Should this worry the particularist? One response is to deny that the only way a moral outlook can be coherent and structured is by resting on a few general moral principles. That the judgments someone makes in different cases can have a shape, hang together, and be consistent only if they are underpinned by general principles is itself a generalist prejudice. Many things can have, or fail to have, a coherent structure: not only scientific theories and mathematical systems, but also narratives, works of art, and human lives. To suppose that moral thought must be modeled on the former rather than the latter is to be, says the particularist, in the grip of the wrong picture.

A different response, not incompatible with this, is to look for a different kind of structure. For example, it might be that we only have reason to do things that either promote the good or benefit someone or something (see McNaughton and Rawling 2008). If so, we can then ask, for instance, whether reason strength is a matter only of the amount of good produced, as the consequentialist would have it, or whether, to take the Rossian line we favor, we can have more reason to benefit those to whom we have special ties (such as duties of fidelity or gratitude) than value alone would dictate.

We have tried to show that ethical intuitionism, in roughly the form advocated by Ross, is defensible. But we have also suggested ways in which it might be modified, if necessary, to accommodate particularist concerns. (See, further, McNaughton and Rawling forthcoming.)[3]

Notes

1 We shall not discuss here "satisficing" consequentialism, which claims you acted rightly if you produced a *sufficient* balance of good over bad. Nor shall we discuss the view that all moral theories can be "consequentialized" (see Portmore 2009).

2 The term does not occur in *The Right and the Good*, and occurs only briefly in Ross's later book *The Foundations of Ethics*.

3 We are grateful to a number of people for helpful comments and discussions both on this topic and on this essay. We would especially like to thank Jonathan Dancy,

Eve Garrard, Brad Hooker, Hugh LaFollette, Ingmar Persson, and Philip Stratton-Lake.

References

Audi, R. (1996) "Intuitionism, Pluralism, and the Foundations of Ethics," in *Moral Knowledge: New Readings in Moral Epistemology*, eds. W. Sinnott-Armstrong and M. Timmons, Oxford: Oxford University Press, pp. 101–36.

Broad, C.D. (1930) *Five Types of Ethical Theory*, London: Routledge.

Copp, D., ed. (2006) *Oxford Handbook of Ethical Theory*, Oxford: Oxford University Press.

Dancy, J. (1998) "Wiggins and Ross," *Utilitas* 10: 281–5.

Dancy, J. (2004) *Ethics without Principles*, Oxford: Clarendon Press.

Darwall, S. (1998) "Under Moore's Spell," *Utilitas* 10: 286–91.

Darwall, S. (2006) "Morality and Practical Reason: A Kantian Approach," in Copp (2006: 282–320).

Hare, R. (1963) *Freedom and Reason*, Oxford: Oxford University Press.

Hooker, B. (2000) *Ideal Code, Real World*, Oxford: Oxford University Press.

Mackie, J. (1977) "The Subjectivity of Values," in *Ethics: Inventing Right and Wrong*, Harmondsworth, UK: Penguin Books, pp. 15–49.

McNaughton, D. and Rawling, P. (2008) "Benefits, Holism, and the Aggregation of Value," in *Utilitarianism: The Aggregation Question*, eds. F. Miller, E. Paul, and A. Paul, Cambridge: Cambridge University Press, pp. 354–74.

McNaughton, D. and Rawling, P. (2011) "The Making/Evidential Reason Distinction," *Analysis* 71 (1): 100–2.

McNaughton, D. and Rawling, P. (forthcoming) "Contours of the Practical Landscape," in *Thinking About Reasons: Essays in Honour of Jonathan Dancy*, eds. D. Bakhurst, B. Hooker, and M. Little, Oxford: Oxford University Press.

Parfit, D. (2011) *On What Matters*, Oxford: Oxford University Press.

Portmore, D. (2009) "Consequentializing," *Philosophy Compass* 4 (2): 329–47.

Ross, W.D. (1930/2002) *The Right and the Good*, Oxford: Clarendon Press.

Ross, W.D. (1939) *The Foundations of Ethics*, Oxford: Clarendon Press.

Stratton-Lake, P. (1999) "Why Externalism Is Not a Problem for Ethical Intuitionists," *Proceedings of the Aristotelian Society* 99: 77–90.

Warnock, G.J. (1967) "Intuitionism," in *Contemporary Moral Philosophy*, New York: St. Martin's Press, pp. 4–17.

Wiggins, D. (1998) "*The Right and the Good* and W.D. Ross's Criticism of Consequentialism," *Utilitas* 10: 261–80.

Further Reading

Audi, R. (1998) "Moderate Intuitionism and the Epistemology of Moral Judgment," *Ethical Theory and Moral Practice* 1: 14–34.

Dancy, J. (1983) "Ethical Particularism and Morally Relevant Properties," *Mind* 92: 530–47.
Dancy, J. (1993) *Moral Reasons*, Oxford: Blackwell, chs. 10–12, pp. 166–233.
Dancy, J. (1993) "An Ethic of *Prima Facie* Duties," in *A Companion to Ethics*, ed. P. Singer, Oxford: Blackwell, pp. 230–40.
Ewing, A.C. (1953) *Ethics*, London: English Universities Press.
Gaut, B. (1993) "Moral Pluralism," *Philosophical Papers* 22: 17–40.
Hooker, B. (1996) "Ross-Style Pluralism vs. Rule-Consequentialism," *Mind* 105: 531–52.
Huemer, M. (2006), *Ethical Intuitionism*, London: Palgrave Macmillan.
McDowell, J. (1979) "Virtue and Reason," *The Monist* 62: 331–50.
McNaughton, D. (1988) *Moral Vision*, Oxford: Blackwell, chs. 11, 13, pp. 163–81 and 190–205.
McNaughton, D. (1996) "An Unconnected Heap of Duties?" *Philosophical Quarterly* 46: 433–47.
McNaughton, D. and Rawling, P. (2000) "Unprincipled Ethics," in *Moral Particularism*, eds. B. Hooker and M. Little, Oxford: Oxford University Press, pp. 256–75.
McNaughton, D. and Rawling, P. (2006) "Deontology," in Copp (2006: 424–58).
Nagel, T. (1986) *The View from Nowhere*, Oxford: Oxford University Press, ch. 9, pp. 164–88.
Prichard, H.A. (1968) *Moral Obligation: Essays and Lectures*, ed. J.O. Urmson, Oxford: Clarendon Press.
Shaver, R. (2011) "The Birth of Deontology," in *Underivative Duty: British Moral Philosophy from Sigwick to Ewing*, ed. Thomas Hurka, New York: Oxford University Press, pp. 126–45.
Stratton-Lake, P. (1997) "Can Hooker's Rule-Consequentialist Principle Justify Rossian Prima Facie Duties?" *Mind* 106: 751–8.
Stratton-Lake, P. (2002) "Introduction," in Ross (1930/2002: ix–lviii).
Stratton-Lake, P., ed. (2002) *Ethical Intuitionism: Re-evaluations*, Oxford: Oxford University Press.
Stratton-Lake, P. (2011) "Eliminativism about Derivative Prima Facie Duties," in *Underivative Duty: British Moral Philosophy from Sigwick to Ewing*, ed. Thomas Hurka, New York: Oxford University Press, pp. 146–65.

Chapter 14

Kantianism

Thomas E. Hill Jr

Among the most basic ideas in Kant's moral philosophy are these: that moral philosophers must use an a priori method, that moral duties are categorical imperatives, and that moral agency presupposes autonomy of the will. In the second section of his *Groundwork of the Metaphysics of Morals*, Kant develops each of these ideas in an argument for his central thesis that the idea that we have moral duties presupposes that we are rational agents with autonomy (Kant 1964; citations to the standard Prussian Academy edition, volume 4 will be included as bracketed numbers). The conclusion and each step of the argument remain controversial. Kant's admirers usually see here a great advance in moral theory, but critics often find Kant's contentions obscure and implausible.

When a philosopher inspires such extremes of admiration and disdain as Kant does in his ethical writings, we may well ask ourselves whether Kant's friends and his critics are focusing their attention on the same ideas. Elementary misunderstandings of Kant's ethics are common, and serious Kant scholars often disagree about interpretations. Insightful core ideas may be dismissed or ignored because they are conflated with more radical, controversial ideas. My aim, then, is to do some much needed sorting among the doctrines attributed to Kant. What is central, and what is peripheral? What is commonplace, and what is radical? Which assertions are preliminary starting points, and which are more remote conclusions? Considering these questions is necessary for a balanced assessment of the strengths and weaknesses of Kant's ethics.

In this essay I comment in turn on each of the major themes mentioned above, trying to separate the more widely appealing core points from the more

The Blackwell Guide to Ethical Theory, Second Edition. Edited by Hugh LaFollette and Ingmar Persson.

controversial. The modest version of each basic theme, I suggest, leads naturally to the next. Together the steps reflect a Kantian line of reflection for his contention that analysis reveals that the idea that we have moral duties presupposes the idea that we are rational agents with autonomy. To preview, my main suggestions will be these:

(1) Kant's insistence on an a priori method, in its modest version, stems in large part from his belief that moral theory should begin with an analysis of the idea of a moral requirement (duty). Despite his strong rhetoric about setting aside everything empirical, Kant's main point was that empirical methods are unsuitable for analysis of moral concepts and defense of basic principles of rational choice. The reason that Kant insisted on an a priori method was not that he believed in rational intuition of moral truths, opposed naturalistic explanations, assumed that duties are imposed by noumenal will, or thought that empirical facts are irrelevant to moral decisions.

(2) Kant thought that analysis of the ordinary idea of duty showed that we regard duties as categorical imperatives. That is, when we suppose that we have a duty we are thereby supposing that we have sufficient (overriding) reason to act accordingly and not just because doing so furthers our (desire-based) personal ends. The modest point here is not that duties must always be experienced as unwelcome demands that must be fulfilled from a sense of constraint. Kant's point is also independent of his dubious view that substantive principles regarding lying, obedience to law, sexual purity, and so on, are exceptionless and applicable in the same way across all times and places.

(3) The analysis of duty is for Kant merely a step on the way to the conclusion that in thinking of ourselves as having moral duties we must think of ourselves as rational agents with autonomy of the will. The basic point is that in order to be a moral agent, with duties, one must be able to understand and be moved by the sort of reasons that categorical imperatives claim we have. Categorical imperatives are addressed to deliberating rational agents presumed able to follow reasons independent of their concern for happiness and personal ends. To think that we can guide our decisions by such noninstrumental reasons, we must conceive ourselves as agents that implicitly acknowledge and respect the noninstrumental rational standards presupposed by categorical imperatives. As moral agents we might not always live up to the standards that we acknowledge, but our capacity to follow them presupposes that we accept them as rational grounds for our decisions and judgments. More controversially, in regarding our duties as categorical imperatives we presuppose that our disposition to judge our conduct by these basic standards is a constitutive feature of being moral agents, and not something we do because of a prior commitment to following external authorities, tradition, or common sentiments. In a sense, then, particular duties can be understood as requirements that rational agents impose on themselves, and following them is a way of being self-governing.

The A Priori Method in Moral Philosophy

Kant repeatedly emphasizes in the *Groundwork*, and elsewhere, that we cannot find answers to the fundamental questions of moral philosophy by empirical methods (Kant 1964: 74–80 [406–12], 92–4 [425–7]). To gain a theoretical understanding of nature we must rely on experience. We must use empirical concepts as well as some basic categories of thought. Ordinary, commonsense knowledge of what there is, how things work, and what is needed to achieve our goals must also rely on experience. But moral philosophy, Kant insists, is not an empirical science, and its conclusions are not simply inferences from observations of human behavior, emotional responses, and social practices. Rather, to address the basic questions of moral philosophy, according to Kant, we must use an a priori method that does not base its conclusions on what we learn from experience. Kant rejects many of the prominent moral theories of his day (e.g., British "moral sense" theories) because they treat moral questions as if they were empirical questions. He rejects, for example, Francis Hutcheson's view that moral goodness is a natural property of actions that causes human beings to feel approbation (Schneewind 1990: 503–24). On this view, the answer to "Which acts are morally good?" would be discoverable by observing what sorts of acts human beings tend to approve. Kant criticizes other theories for mixing empirical and a priori arguments in discussions of basic issues that, he thinks, should be approached in a purely a priori manner. For example, Kant strongly disapproves of moral philosophies that argue that helping those in need is right and reasonable *because* experience shows that charitable people tend to be happier than uncharitable people.

Why begin moral philosophy by an a priori investigation instead of empirical studies? The explanation, I think, concerns Kant's understanding of what the basic questions of ethics are. In the *Groundwork*, he describes his task as seeking out and establishing the supreme principle of morality (Kant 1964: 60 [392]). Judging by how Kant then proceeds to argue, it seems that "seeking out" the supreme principle is a matter of articulating an abstract, basic and comprehensive principle that can be shown to be a deep presupposition in ordinary moral thinking. "Establishing" the principle, I take it, is the further task of showing that the principle is rational to accept and follow. In addressing the first task, Kant begins, provisionally, by assuming some very general moral ideas that he takes to be widely accepted – in fact part of ordinary rational knowledge of morality. These assumptions include the special value of a good will and the idea of duty as more than prudence and efficiency in pursuing one's ends. That these are only assumed provisionally is shown by the fact that, even at the end of the second section, Kant forcefully reminds us that his "analytic" mode of argument has not proved we really have moral duties (Kant 1964: 112 [444–5], 107–8 [440–1, 114–15 [446–7]). Instead, it only serves to reveal presuppositions of the common moral idea that we have duties. For all we know at this point, morality might be an illusion.

Despite this disclaimer, the results that Kant claims to reach by the analytic method are significant: common moral belief presupposes that the several formulas of the Categorical Imperative are morally fundamental, that rationality is not exclusively instrumental, and that moral agents are to be seen as legislators of moral laws as well as subject to them. These particular conclusions, however, are supposed *results* of the a priori method of analysis, not assumptions used to justify the method. Other philosophers might radically disagree with Kant's results but still see the value of his analytic approach.

Kant's main idea is simple and familiar in philosophy. We make use of moral concepts, some of which seem pervasive and essential features of our moral thinking and discourse, even when we disagree in our particular judgments. By reflecting on the meaning, implications, and presuppositions of these concepts, we may be able to understand them, and ourselves, better. To say that the process of reflection is a priori is not to imply that it could be done by hypothetical persons with no empirical concepts or experience of life. It is just to say that we are examining our ideas in a rational, reflective way, looking for their structure and presuppositions. The aim here is not to explain the causes or effects of behavior that seems to be guided by moral ideas but only to gain a clearer grasp of the content and implications of those ideas themselves. Experiments, surveys, and comparative studies of different cultures can be valuable for many purposes, but they do not serve the philosophical purpose that Kant's analytical method was meant to address.

There was another important reason why Kant wanted moral philosophy to begin with an a priori method. This stems from his conviction that believing that we are under moral obligation entails believing that we are subject to a rational requirement of a special sort (a "command of reason"). This conviction was embedded in a long tradition, and Kant thought that it was part of ordinary understanding of morality. The problem is that we can question whether the *apparent* rationality of moral demands is an illusion. In fact, reading the British moralists Hutcheson, Hume, and others would naturally raise doubts in those (like Kant) who were deeply influenced by the natural law tradition. Such doubts, Kant thought, call for a response: an effort to vindicate the apparent (and commonly believed) assumption that moral principles express requirements that we would be irrational to disregard (Kant 1964: 114–31 [446–63]). A positive response to the doubts would supplement (and build on) the analytical argument mentioned above with a further argument that we really have reason-based duties, or at least that it is necessary to presuppose this for practical purposes. To do so would be to show that morality is not a mere illusion. Like the task of analysis, this task, which Kant undertakes in the notoriously difficult third section of the *Groundwork*, is again not one that could be achieved by empirical investigations. The problem is to establish that guiding one's life by certain principles is *rationally necessary*; that one always has *sufficient reason* to do so.

Even if (contrary to Kant) there are only prudential reasons for following moral principles, to show that following them is always rational is not *simply* a matter

of collecting empirical data on the effects of various behavior patterns. One would also need to argue that we always have *sufficient reason* to do what most effectively promotes the effects deemed "prudent," and this is a contested philosophical thesis that is not itself subject to empirical proof (as even most non-Kantians would agree). But the inadequacy of using an empirical method alone becomes even more evident for those who grant Kant's thesis that morality imposes categorical imperatives (Kant 1964: 82–8 [414–20]). According to this, moral principles are rationally necessary to follow, but their rational necessity is not merely prudential or based on hypothetical imperatives. This means (at least) that the reason for following moral principles cannot be simply that doing so serves to promote one's happiness or individual ends. Thus, the rationality of following moral principles could not be established by showing empirically that they are good guides to happiness or means that serve well our particular purposes. For not only is the idea of rationality a normative one (the previous point), but also the sort of sufficient reason that needs to be defended is more than the (empirically discernible) efficacy of our actions in achieving our ends.

This is not the place to review and assess Kant's actual argument in defense of his idea that moral requirements are *rationally necessary* to follow and even *categorically* so. And this assessment, fortunately, is not necessary for present purposes. The need that Kant saw for an a priori method, at least in parts of ethics, can be seen in the *problems* he posed, independently of his particular solutions. The essential point is that *if* we understand moral demands as saying to us that it is *unreasonable* not to do what is demanded, then we want some explanation and defense, especially once philosophical doubts have been raised. All the more, if we understand moral demands as purporting to tell us what is *categorically* rational to do, then we may question whether morality's claim to be categorically rational is defensible. If, like most contemporary philosophers, we understand that claims about what is *reasonable*, *rational*, supported by *reasons*, and so on, are irreducibly evaluative, practical claims, then it becomes clear that the problems cannot be resolved by empirical investigation alone. The problems may prove to be irresolvable, or perhaps even pseudoproblems (as Humeans think), but at least we can understand why Kant and others believe that any search for resolutions must start with rational, a priori reflection.

Now that we have uncovered Kant's rationale for thinking that we must employ an a priori method, we can respond to some common objections and clarify certain misconceptions about the method.

(1) One misunderstanding that might lead readers to be skeptical of Kant's methodology stems from the thought that the alternative to empirical methods in moral theory is appeal to rational intuition or rationalistic theological arguments. Hume's famous objections to deriving "moral distinctions" from "reason" seem primarily aimed at views of this type. If turned against Kant, however, objections to rational intuition and theological ethics would miss their mark, for Kant agrees with Hume in rejecting rational intuitionism and theology as the basis of ethics.

Like Hume, Kant holds that the traditional a priori arguments for the existence of God are inadequate, that morality cannot be based in theology, and that reason is not an intuitive power that "sees" independent moral facts. (Kant does not deny that there is "knowledge" of moral principles and that there are "objective" moral values, but moral validity is determined by, and so not independent of, what rational agents with autonomy could (or would) accept.)

(2) Some moral theorists, past and present, see their main task as explaining moral phenomena as a part of the natural world. It seems obvious that we raise moral questions, praise and blame in moral terms, experience moral feelings (e.g., guilt, indignation), and are sometimes moved by our moral beliefs. Many philosophers committed to understanding the world, so far as possible, in naturalistic terms accept the challenge of trying to explain moral phenomena (behavior, feelings, etc.) without appeal to occult, theological, or other "nonnatural" entities. The methodology needed for this project, it seems widely agreed, is empirical, at least in a broad sense. When we turn to Kant's moral philosophy we find that not only does he use terminology (e.g., the will, autonomy, intelligible world) that is outside what most naturalists consider their domain, he also even insists that these moral terms cannot be understood entirely in naturalistic terms. Clearly his moral theory is not a successful fulfillment of the naturalists' project, and may even seem to reflect contempt for such a project. Thus an objection to Kant's a priori method might be grounded in the thought that it is a method that cannot successfully carry out the project that naturalists consider most important and may even show contempt for it.

It is true, of course, that Kant's moral philosophy is not an attempt to contribute to the naturalists' project, but this does not mean that he would regard it as an unfruitful or unimportant task for empirically oriented scientists and philosophers to undertake. Although Kant insists that the a priori tasks in moral theory must be undertaken first, he often refers to "practical anthropology" as empirical work that should follow and supplement basic moral theory (Kant 1964: 55–6 [387–8]; Kant 1997a). What he had in mind (and attempted rather casually and unsystematically) was not the full naturalists' project, but his theory of knowledge is friendly to that project, at least if no more is claimed for its results than can be validly inferred from experience. Kant is committed to the position (which in fact he believed that he had proved) that all phenomena are in principle explicable by empirical, natural laws. So, although he thought that for practical purposes we must employ normative ideas that are not reducible to empirical propositions, anything that can count as observable phenomena associated with moral practices must (in principle) be amenable to empirical study and understanding. And, although he denied that empirical science can establish moral truths or vindicate their rational claim on us, his theory of knowledge allows (indeed insists) that all the *observable facts* associated with moral and immoral acts can be studied and (in principle) comprehended from an empirical perspective. This is distinct from the practical perspective we must take up when we deliberate and evaluate acts (see

Allison 1991). Each perspective has its legitimate and necessary use, and limits. So, although Kant thinks the basic issues of moral philosophy cannot be answered by empirical methods, he should happily encourage a naturalist's ambition to understand the phenomena associated with moral activity *so far as possible* in naturalistic terms through empirical investigations.

(3) Again, some critics familiar with Kant's philosophy as a whole may suppose that Kant's insistence on an a priori method is based on his controversial idea that we must think of moral agents not only in empirical terms but also under the idea of free rational agency. This involves thinking of them as belonging to an "intelligible world" that cannot be understood in the terms of empirical science (Kant 1964: 118–21 [450–3]). Hence one might suspect that Kant thought an a priori method of investigation in ethics is necessary because moral agents, as such, are not beings that we can comprehend empirically. But I think that this is a mistake, and in fact it gets the order of Kant's thought backwards. As we have seen, there are simpler and less controversial explanations for Kant's insistence on the a priori method. In fact he introduces the perspective of an intelligible world into ethics not as an initial assumption but rather as a point to which he believes his analysis of common moral knowledge finally drives him. Analysis of the idea of duty shows that it presupposes the idea of rational agents with autonomy, and this idea, he argues, can be squared with his earlier conclusions about empirical knowledge only if we think of these agents as "intelligible" or noumenal beings.[1] Many philosophers who find Kant convincing at the earlier stages dissent from this last stage of the argument. There is no doubt that Kant thought it an important part of his systematic moral theory, but it is not a beginning assumption used to justify his methodology. Rather, it is a final theoretical point to which (Kant thought) his particular a priori argument (not the method itself) drives us. In short, his controversial views about the ultimate "Idea" of moral agents to which philosophical reflection forces us is not presupposed in the modest methodological procedures with which he begins.

(4) Finally, there is a persistent objection that, I suspect, rests partly on misunderstanding but partly on Kant's tendency to overstate his insights. The objection proceeds as follows. First we note that the reasons we give for thinking that acts are right or wrong are typically empirical facts, for example, "That will kill him," "You intentionally deceived him," "She saved your life and needs help now," "No society could survive if it tolerated that." Then we also note that most morally sensitive persons realize that the acts picked out by simple descriptions (e.g., "killing," "deceiving") may be wrong in one situation but right in another, depending on the empirical facts of the case. So a method that excluded empirical information, it seems, will not even consider facts that are crucial to determining what is right and what is wrong. Moral decisions must be made in a complex and richly diverse world, and so it seems foolish to suppose that we can discern what is right without knowing accurately and in detail (and so empirically) what this world is like and where we stand in it at the moment.

The objection would be appropriate and (I think) devastating if directed against a moral theorist who claimed that pure reason alone can discern what we ought to do in each situation. But few, if any, today make such a claim, and certainly Kant did not. Those who agree with Kant that some fundamental moral principles can be vindicated through the use of reason are well aware that we need empirical knowledge to apply these principles to our current circumstances. We need to judge whether and how moral principles are relevant, and this requires understanding based on experience. For example, that we should treat all persons with respect, Kant thought, is an ideal norm, not something empirical science or ordinary experience can establish; but, of course, respect and disrespect are expressed in a wide variety of ways that we learn only with experience in different cultural contexts (Kant 1996: 209–13 [462–8]). Kant does not deny that we (rightly) cite facts in explaining the reasons why some particular act is morally required or forbidden; he merely agrees with Hume that empirical facts *alone* do not establish any "ought"-claim. Kant was indeed extremely rigoristic by not allowing that familiar moral principles (e.g., about lying) need to be qualified, but his rigidity on these matters cannot be blamed on his rejection of empirical methods for the basic issues in moral theory. Notoriously, Kant endorses some principles in an absolute, unqualified form, and most of us will agree that inflexible adherence to such rules is an oversimple response to complex moral problems. His extreme stands on lying, revolution, and sexual practices, however, do not follow from his thesis that moral philosophy should *begin* with a priori methods, for example, of analysis (Kant 1996: 176–7 [422–3], 96–7 [320], 178–7 [424–5]; Kant 1993). The problem, rather, lies in his thinking that rigid opposition to lying (etc.) is required by the Categorical Imperative.

There remains serious controversy, however, on two related points. First, many philosophers would deny that an a priori use of reason can establish even one basic moral principle. This objection comes not only from those who think that empirical methods can establish moral principles, but also from those who think that moral principles cannot be established by any method because they have no objective standing. This is a perennial controversy, but it is about the *results* that can be established by an a priori method rather than about the value of the method in general. Second, even those who side with Kant on the first point may reasonably worry that Kant himself tries to make *too much* of ethics independent of empirical knowledge. It is one thing, they may say, to suppose that some quite abstract, formal principles can be discovered and defended by an a priori method, but quite another (and more dubious) thing to exclude empirical facts when taking up other tasks of moral philosophy. For example, if moral philosophers, following Kant and Alan Donagan, want to try to work out a system of universally valid moral principles about substantive matters (such as lying, obedience to law, punishment, charity), then it seems only reasonable to expect that the construction must take into account our (limited) empirical knowledge about the human condition in general and about the diversity of contexts to which putative universal principles must be applied. It is still a matter of dispute how much empirical

information Kant intended to exclude when he took up this project in *The Metaphysics of Morals*. His arguments often presuppose facts that could only be known empirically, but they also often raise the suspicion that his determination not to rely on empirical evidence has led to unwarranted rigidity and overgeneralization. These worries and controversies cannot be lightly dismissed, but they do not call into question Kant's main reasons for adopting an a priori method for the basic issues in moral philosophy.

Categorical and Hypothetical Imperatives

The vocabulary and tone of Kant's writing about morality are disturbing to many readers, especially when they contrast these with the ethical works of Hume and Aristotle. A good example is Kant's contention that there are *categorical imperatives* of morality. Kant focuses attention on what we morally *must* do, what is *necessary*, a *command* of reason, a *constraint* rather than an aid in the pursuit of happiness (Kant 1964: 82–8 [414–21]). We are easily reminded of angry parents who tell us, in stern imperative tones, "Do it at once, whether you want to or not." So viewed, morality can seem to be dictatorial, not intrinsically appealing or personally fulfilling. Moreover, since Kant tells us that categorical imperatives are unconditional, absolute, and apodictic as opposed to mere prudential "counsels," it is natural to assume that this means that moral rules are inflexible and admit of no exceptions. This assumption may seem confirmed when we read Kant's vigorous denial that we may tell a lie to save a friend from murder and his insistence that we must obey the law even if it is imposed by a tyrant (Kant 1993; Kant 1996:127–33 [316–23] and 176 [371]). Categorical imperatives then seem like demands that we must obey with the attitude of a dutiful soldier following orders, respecting the authority of law without regard to anything else.

Kant's moral theory no doubt contains features with which many ordinary readers, as well as opposing moral theorists, will disagree, but making some distinctions helps us to identify some possible misunderstandings and to sort the more controversial from the less controversial Kantian themes. There may remain disputes about both interpretation and plausibility, but I think that some core ideas that are manifestly at least part of Kant's thought are also quite widely accepted. Three questions, in particular, need to be considered: (1) Are categorical imperatives to be seen as disagreeable orders from an alien power with whom we cannot identify – mere pressures that we see no good reason to follow apart from possible rewards and punishments? (2) Are moral principles, as categorical imperatives, necessarily inflexible and exceptionless? (3) Is a motivating respect for principles that are categorical imperatives necessarily a sense of constraint rather than concern for the good of others?

Despite what one might initially suppose, Kant's basic position on each of these questions, I think, is quite compatible with common opinion (among philosophers

and nonphilosophers alike). This is not to deny, however, that Kant accepts some further related ideas that remain more controversial. Let us begin with what I take to be the core idea that moral duties are categorical imperatives, and then we can return to the three questions just mentioned.

Kant's remarks about categorical imperatives can be confusing because although he explicitly says that there can only be one Categorical Imperative he repeatedly writes as though there are many. Kant lists several "formulas" of the Categorical Imperative, which he says are "at bottom the same," but he also refers to more specific principles, such as "Do not lie" and "Punish all and only the guilty" as categorical imperatives (Kant 1964: 88–104 [420–37]; Kant 1996: 14 [221] and 105 [331]). No doubt he had in mind a primary (or strict) sense of the term when he was writing as if there were only one Categorical Imperative, but he then helped himself to a secondary (or less strict) sense of the term when writing about further principles that (he believed) were warranted by "the Categorical Imperative" (in the strict sense).[3] On this hypothesis, the discrepancy (from singular to plural) becomes harmless, even though there remain questions in various contexts about which sense he had in mind.

Categorical imperatives (in both senses) are *imperatives*, which Kant calls "commands of reason." All imperatives express the idea that something ought to be done, either because it is good in itself or because it is good as a means to an end that is in some way valuable. Through the idea of "ought" they express a relation ("necessitation") between what is rational to do (an "objective principle") and the not so perfectly rational choosers ("imperfect wills") that can do what is rational but might not (Kant 1964: 81–4 [412–17] and Kant 1964: 69n [401]). So, in other words, imperatives say (truly) that we have good reason to do something even while acknowledging (implicitly) that we might in fact not do it. This applies to "hypothetical imperatives," for example, "one ought to exercise if one aims to be strong," as well as to "categorical imperatives," for example, "one ought to treat human beings with respect."

What, then, makes an imperative *categorical*? For both the primary and secondary senses, the core idea is that the reasons for following a "categorical imperative" are not merely that doing so will promote the ends that one happens to have, such as becoming rich or (more generally) being happy. Following categorical imperatives may often promote our personal ends, but it may not always do so. Making us happy and helping us get what we want are not what makes moral principles *categorical* imperatives; they are rational to follow, even if doing so does not make us happy or promote our personal ends. They express the idea that it is good and rational to act as they prescribe, but, unlike hypothetical imperatives, they do not simply say what is good to do as a means to getting or achieving what we want.

Furthermore, as Kant uses the term, categorical imperatives do not merely say that we have we have *some* reason to do what they prescribe. They assert that we have sufficient reason, *overriding* other considerations.[2] We always ought to follow categorical imperatives, even if they conflict with what we otherwise would have reason to do based on self-interest and our personal projects. So categorical

imperatives do not simply give us "some" reason to act; they give us sufficient reasons, all things considered – reasons that override other considerations. This point, however, should not be confused with the idea that moral rules are always specific, simple, and inflexible, allowing no exceptions or variation for extraordinary circumstances. Kant himself did insist on *some* moral principles (e.g., against lying) in this rigid form, but nothing in the core idea of categorical imperatives prevents them from being vastly complex and justifiably filled with qualifications ("unless," "so long as," "but only if"). Moreover, as Kant says, some ethical principles only say that we ought to adopt certain indeterminate ends (e.g., the happiness of others), without specifying exactly what, or how much, one must do to promote the ends (Kant 1996: 147–56 [382–94]). These too are supposed to be categorical imperatives, for they say we must, for overriding reasons, adopt the prescribed ends, whether or not doing so promotes our happiness and personal projects. So categorical imperatives do not have to be inflexible, rigoristic rules of conduct. In labeling an "ought"-judgment as a "categorical imperative" we express the belief that it is an all-things-considered, overriding moral requirement, backed by reasons not entirely dependent on what serves to promote the ends we happen to have. The requirement could be simple and sweepingly general (as Kant regarded "Never lie"), but it could be vastly complex and qualified. We should not confuse issues about the scope and complexity of moral principles with issues about the sort of reason we have to follow them. Kant's claim that we are under categorical imperatives is addressed to the latter.

Beyond these core ideas, Kant held that the one "Categorical Imperative" (in the strict sense) is an unconditional and unqualified requirement of reason, applicable in all human conditions and implicitly acknowledged in common moral judgments. He expressed this Categorical Imperative in several formulas, which are supposed to articulate the supreme moral principle in somewhat different ways. In sum, the formulas (and letters commonly used as labels) are: (1)(i) The formula of universal law (FUL): "*Act only on that maxim through which you can at the same time will that it should become universal law*" (Kant 1964: 88 [421]); (ii) A variation – the universal law of nature formula (FULN): "*Act as if the maxim of your action were to become through your will a universal law of nature*" (Kant 1964: 89 [421]). (2) The formula of humanity (FHE): "*Act in such a way that you treat humanity, whether in your own person or in the person of any other, never simply as a means, but always at the same time as an end*" (Kant 1964: 96 [429]). (3)(i) The formula of autonomy (FA): "The supreme condition of the will's conformity with universal practical reason [is] the Idea *of the will of every rational being as a will which makes universal law*" (Kant 1964: 98 [431]); (ii) A variation – the kingdom of ends formula (FKE): "A rational being must always regard himself as making laws in a kingdom of ends which is possible through freedom of the will . . ." (Kant 1964: 101 [434]) and "All maxims as proceeding from our own making of law ought to harmonize with a possible kingdom of ends as a kingdom of nature" (Kant 1964: 104 [436]). (Common interpretations are summarized in Hill 2006, but my concern here is with other basic Kantian themes.)

Unlike the principle behind instrumental reasons, which we might call "the Hypothetical Imperative," the formulas of the Categorical Imperative do not simply prescribe taking the necessary means to desired ends. They can be established as rationally necessary, Kant thought, without reliance on empirical studies of human nature. It expresses what our own reason, independently of inclination, requires of us, and so we cannot help but acknowledge its authority (even when we fail to meet its requirements). Kant seems at times also to believe that more specific principles (e.g., about lying, obeying the law, and sexual practices) are derivative categorical imperatives, shown by the basic Categorical Imperative to be unconditionally required in *all human conditions*, without exception. These ideas are understandably more controversial than the basic points we have been discussing.

Kant's core idea, however, remains just that moral duties impose categorical imperatives in the sense that we have sufficient, overriding reason to fulfill our moral duties, independently of whether doing so will promote our own happiness or serve our individual ends. Even this core idea is rejected by those philosophers who insist that practical rationality is always nothing but taking efficient means to desired ends, but Kant's view, I suspect, is closer than theirs to ordinary moral opinion and most of Western tradition in moral theory.[4] We think, for example, that Hitler was wrong and *unreasonable* to kill millions of European Jews and this was not just because it was a poor means for him to get what he most wanted. The moral prohibition on murdering people, it is commonly thought, should override personal ambitions; so Hitler had sufficient reason not to do the killing, even though he wanted to.

Now let us return to our earlier questions.

(1) It should be clear that categorical imperatives are not to be viewed as orders from an alleged alien authority. Unlike commands from parents, military superiors, and legal authorities, they are conceived as expressing "objective principles"; that is, principles that anyone in the context would follow if sufficiently guided by reason. They are supposed to tell us what is good in itself to do, not what someone demands that we do.[5] As discussed more fully in the next section of this essay, "Autonomy of Moral Agents," a key Kantian doctrine is that basic moral requirements are laws we legislate to ourselves as rational persons with autonomy. We are not morally bound by any alleged requirement unless it is backed by principles that we can recognize as what we ourselves, as rational, self-governing persons, will for ourselves and others. There are various ways of understanding this, but all clearly rule out the idea that categorical imperatives are imposed by alien authorities and give us reasons only by threats of punishment or promise of rewards. The authority of moral principles is, as it were, the authority of our own reason, our best judgments, all considered, as to what we ought to do. Moral reasons are *our* reasons; they guide us, rather than goad us (Falk 1986). What they require need not be unpleasant or disagreeable at all; but even when it is, we cannot pursue other projects in disregard of them without going against

our own best judgment, suffering conflict of will, and inviting self-contempt. These implications of Kant's idea of moral autonomy may be doubted, but at least they make clear that Kantian categorical imperatives would be grossly misunderstood if they were seen as commands of some alleged "authority" independent of our own reason.

(2) It should also be clear that substantive categorical imperatives need not be simple, exceptionless rules, like "Never lie." As noted earlier, Kant himself believed that there are such absolute rules, but this dubious belief does not follow from the concept of a categorical imperative. What follows is that, no matter how richly complex and filled with "unless" and "so long as" clauses, a categorical imperative should always be respected, not subordinated to other considerations. To call a specified requirement a "categorical imperative" is to make a summary judgment, saying that, all things considered, reason requires a certain course of action. If we believe that a principle states merely a morally relevant consideration, then it should not be called a categorical imperative; for that label is appropriate only when all relevant factors have been taken into account and an all-things-considered conclusion on a particular act or act type has been reached. We can say, trivially, "categorical imperatives must be obeyed, no matter what," because the claim is implicit in what is meant by a categorical imperative. Again, however, nothing follows about the complexity and scope of the principles that summarize our reasonable all-things-considered moral judgments about lying, revolution, sex, promises, and so on – that is, the principles we might take to be categorical imperatives. In short, we should not confuse two distinct questions: (i) How much (if at all) should moral principles about lying, killing, obeying the law, and so on, be qualified by explicit or implicit exceptions? and (ii) Are moral principles *categorical imperatives*? The core issue for the second question is whether moral principles, no matter how many or few qualifications they contain, are overridingly rational to follow and not simply because doing so promotes the personal ends of the agent.

(3) Finally, a categorical imperative is not something we must follow from a sense of constraint. We do not need to grit our teeth and focus on the requirement as a "command," to which we are "bound" and "subject." We can often, and should, fulfill our moral responsibilities with our mind focused on the good we can do, rather than our own goodness or need to submit to authoritative commands. This requires some explanation.

To be sure, Kant does imply that in general we are not only authors of moral laws but *subject* to them (Kant 1964: 98–102 [431–4]). As *imperatives* they express a relation of *necessitation* between our imperfect wills and objective principles, that is, the principles that we would follow invariably if we acted in a fully rational way (Kant 1964: 80–1 [413]). Moreover, Kant says that conforming to duty has "moral worth" only if done "from duty" (Kant 1964: 65–7 [397–9]). But none of this, I think, implies that we are *always or typically* averse to doing what we should or that we need to feel "constrained" in order to do it. If we are

in fact reluctant to do what we should, then the thought that doing so is an *imperative* to which we are *subject* may serve to move us (or not); but the thought is not essential, I think, to the idea of governing ourselves by principles that are categorical imperatives. Kant did tend to suppose that self-interest is such a strong motive that recognition of the moral law inevitably causes in us feelings of "respect," and he describes this respect as a partly painful feeling, akin to fear, a sense of our "self-conceit" being humbled by recognition that what is morally required is not always what we most want to do (Kant 1997b: 62–75 [71–89]). This dark, and perhaps overly pessimistic, view of human psychology, however, is not an implication of the core idea that moral requirements are categorical imperatives. That idea is about the sort of reasons that favor acting as we morally should, leaving open whether on particular occasions acting for those reasons will be experienced as being constrained or obedient to authority.

An important point to note here is that all "imperatives" have two sides, as it were. They express "objective principles," or rational principles that even a "holy will" would follow, and yet they do so in a form ("ought") that also conveys the idea that imperfect wills, that is, those who do not automatically follow them with Godlike regularity, are *bound* to obey them, *must* do so, and feel *constrained* when tempted to do otherwise.[6] When we consider our thoughts and feelings in fulfilling our various duties, then, there are several possibilities.

First, we might desire to do something incompatible with what we ought to do but nevertheless understand and respect the reasons behind the moral principle. Here it seems natural to suppose that we are moved both by the moral reasons and by a sense of being under appropriate constraints. Suppose, for example, you are asked to testify in a legal case, and telling the truth will prove embarrassing to you and your friend, but you recognize and respect the moral reasons for obeying the law and testifying truthfully, and you conclude that, all things considered, this is your duty. If you tell the truth, you do so because you respect the good reasons for doing so but also with a sense of being constrained to do so contrary to your wishes. Or, better, you have had to *constrain yourself* to act on principle rather than inclination. This is the sort of case, I think, that Kant most often highlights.

Second, we might desire to do something incompatible with what we accept as duty but without understanding or even considering the good reasons for accepting it as duty. We might have just relied on common opinion or the authority of another person. Here we would fulfill the duty with a sense of constraint but not from respect for the reasons behind it. Insofar as we have really accepted the opinion that we have the moral duty, then by Kant's analysis we must suppose that *there are* sufficient, overriding reasons but we are not aware of them and so cannot be moved by them. An example might be a person who restrains sexual impulses according to commonly accepted opinions about what is permissible but who never considers why those restraints are required.

Third, now that we see that the elements of Kant's paradigm case (i.e., the first case above) are separable, we can consider another possibility. That is, we might

recognize and respect the reasons for a particular moral requirement but have no inclination or reason to act otherwise. The suggestion is not that we *never* think of duty, but just that in the case at hand there is no need for constraint because nothing even prompts the thought of not doing the right thing. Suppose, for example, your child is badly cut from a fall and needs hospital treatment immediately. In fact, as you would agree if asked, it is your duty to take the child to the hospital, but the constraints and imperatives of moral duty are not at all what is on your mind. Nor are you thinking "The child is *mine* and so I must help." Your love draws your attention to the need of Ken or Leah, the individual person in front of you perceived concretely.[7] The life and interests of *this* child are so clear and vivid that abstract thoughts about all human beings' reasons for helping other human beings are not what is on your mind. But, still, what primarily moves you in the particular context are features of it that would give anyone reason to act similarly in relevantly similar cases. It is not that the child is named "Ken" or "Leah," or anything else in particular. It is not primarily, certainly not only, that the child shares your genes or has lived with you for several years. You are moved by a direct concern for the life and vital interests of the real person, and these are the very sort of reasons about which moral principles speak in more general terms. Your reasons, I would say, are moral reasons insofar as they manifest in the particular case the sort of attitude that the more abstract principles of humanity and beneficence call for. The motivation does not fail to be respect for moral reasons just because it was not the result of a deduction from abstract moral generalizations to particular cases. It seems, then, that we can be moved by the relevant moral reasons without experiencing them as constraints and without even thinking of them in the form of abstract generalizations. If so, even imperfect moral agents, like us, who often experience moral requirements as constraints, need not always do so. In at least some circumstances we can act as categorical imperatives prescribe, responding directly to the reasons behind them, without experiencing them as constraints or even thinking of them abstractly as duties.

Would Kant count these acts as "from duty" and so "morally worthy"? The answer is not entirely clear because the idea of duty includes the two elements (moral reasons and constraint) that can work separately in ways that Kant did not discuss. Even if Kant assumed that the sense of being (self-) constrained is an essential part of acting "from duty," a reasonable extension of Kant's view would, I think, grant that the crucial feature of morally worthy acts is that they manifest responsiveness to the sort of basic reasons that underlie moral principles.

Autonomy of Moral Agents

Kant argued, still by an analytical method, that there can be only one Categorical Imperative, which he expressed initially in his famous formula of universal law (Kant 1964: 88 [420–1]). In a complex and controversial course of argument, he

contended that this formula expresses essentially the same basic moral idea as his later formulas, including the formula of autonomy (Kant 1964: 88–100, 104–8 [420–40]). According to this formula, we must act under the idea that moral agents legislate or will for themselves universal laws, as rational beings, independently of their particular desires as sensuous human beings. Thinking of ourselves as under the Categorical Imperative, then, requires thinking of ourselves as rational agents with what Kant calls autonomy of the will. Thus, assuming Kant's analysis of the idea of moral duty as the idea of being subject to categorical imperatives and so bound by the Categorical Imperative, believing that we have duties commits us to a conception of ourselves as rational agents with autonomy. Now there is much in this whole argument that for present purposes we can bypass. The core point is Kant's thought that we must attribute at least a modest sort of autonomy to moral agents because we think of them as having the capacities and dispositions to guide their decisions by categorical imperatives. Kant also affirms a more robust, and controversial, conception of autonomy, in line with his stronger claims about the Categorical Imperative, but let us begin with the more modest idea.

What sort of agents could be subject to categorical imperatives? All imperatives are rational requirements addressed to those who can fulfill them but might not, and so the agents must be able to follow the rational requirements, recognized as such. That is, they must be disposed to acknowledge and follow them because they are requirements that express good reasons or are based on good reasons. Since being under an imperative implies the possibility of acting against reason, agents subject to categorical imperatives may in fact fail to follow them, and may even act against them; but insofar as we suppose the agents *ought* to follow the imperatives, we must assume that they *can*. Already it is clear, then, that agents subject to categorical imperatives cannot be complete slaves to the impulses and desires of the moment, for that implies inability to regulate conduct by rational reflection, even about future consequences to oneself. At a minimum the agents must be able to act for reasons, reflecting on facts and interests over time. This much is implicit even in the idea that they can follow hypothetical imperatives. Since, however, categorical imperatives are defined as principles rational to follow independently of how well they serve our happiness and particular personal ends, agents subject to them must also be able and disposed to recognize reasons to act beyond those of instrumental rationality. Their deliberations are not restricted to considering what will satisfy their immediate desires, what will make them most happy in the long run, and what will achieve their desires for others. Apart from these considerations, they also acknowledge reasons of another kind, considerations that also other agents, so far as they are rational, accept as reasons and not just because their desires as individuals would be served. Agents subject to categorical imperatives, then, cannot take the fact that they can satisfy a particular desire or interest as sufficient, by itself, to give them a reason to act; for they realize that further reflection, on rational considerations not so tied to their personal concerns, may give them reason to disregard, suppress, or even try to eliminate

that desire or interest. Furthermore, if they judge that, all things considered, these reasons are sufficient to constitute duty, understood as a categorical imperative, they regard them as overriding reasons – determining what they ought to do, despite any inclination not to. Agents conceived in this way have the main elements of a modest Kantian idea of autonomy (for more detail see Hill 1992).

In thinking of agents as having desires but able to reflect to determine whether those desires, all considered, provide good reasons, we are already attributing to them a necessary condition of autonomy. To follow categorical imperatives, however, agents must also be able to acknowledge and act on reasons that are more than requirements to take the means to satisfy their desire-based ends. This is a further feature of Kant's idea of autonomy. When we add that, to follow categorical imperatives, they must respect these special reasons as overriding their desire-based reasons, we have a fuller, but still modest, idea of Kantian autonomy. Some philosophers deny that moral agents must have autonomy even in this limited sense, but the ideas regarding autonomy that draw the most controversy go beyond the basic points mentioned so far.

First, Kant held that moral agents, in a sense, impose moral requirements on themselves. They are *authors* of moral laws as well as *subject* to them. They can be compared to autonomous states, bound to no higher authority, with a power to govern themselves in accord with their own constitution, without needing the approval of any further authority. These metaphorical descriptions may be understood in several ways, but some basic points seem clear. Rational agents with autonomy identify with the perspective from which moral judgments are made so that they see moral requirements not as externally imposed, for example, by cultural norms or divine commands. They cannot, then, knowingly act contrary to their moral beliefs without inner conflict and self-disapproval. When they act from moral principle, they are governing themselves by their own standards; and when they act immorally, they are in conflict with deep commitments essential to them as moral agents. Also, in conceiving of moral agents as "authors" of moral laws, Kant implicitly contrasts his idea of rational autonomy with rational intuitionism. That is, reason does not simply "perceive" moral facts as things that exist independently of the use of reason by moral agents; rather moral agents determine particular moral requirements through reasoning from a basic moral perspective (as if legislating according to values inherent in their constitution).

Second, Kant apparently thought that virtually all sane, competent adult human beings have the characteristics of autonomy that his analysis revealed as essential to moral agency. This, however, is a point of faith beyond what his analytical argument aims to establish. That argument, at best, shows that the idea of a moral agent who acknowledges duties presupposes that such agents are rational and have autonomy. But whether all, or even most, functional adult human beings are moral agents in this sense cannot be settled by conceptual analysis, and Kant, of course, did not undertake any empirical investigations to give evidence for his assumption. In our times, after the Holocaust, it is harder to share Kant's faith that a moral point of view is universally acknowledged as authoritative. Kant tries to make sense

of moral life by offering an abstract model of moral agents with certain essential features; but whether that model fits this or that person, that is, whether they are moral agents in his sense, depends on what we find when we try to employ it. Merely finding examples of sociopaths that fail to be moral agents in Kant's sense, however, does not show that Kant's argument was incorrect or his model valueless. Instead, it would confirm doubts about the common eighteenth-century faith, which Kant shared, that all minimally rational human beings implicitly acknowledge moral standards. Some Kantians will defend Kant on the point; and some critics may argue that Kant's model does not even fit ordinary moral agents. Controversy here is not easily resolved.

Third, Kant held that rational agents with autonomy act from pure practical reason alone. When they act from respect for overriding moral reasons, then, they are not to be understood simply as acting on good (morally approved) *sentiments* as opposed to other desires and inclinations. It is a familiar Kantian theme that they act on principle, where the governing maxim is not of the form "I will do X because, as it happens, X promotes Y, which I want" but, rather, "I will do X, regardless of its effect on what I desire." The claim that we can act from pure practical reason, however, goes beyond these familiar Kantian themes. A sophisticated Humean, for example, might accept those themes but insist that the agent's underlying motive for adopting the maxim of duty is a strong, but "calm," sentiment in favor of so acting. The feelings that move us are not always reflected in the maxims we use to guide and explain our conduct. Even Kant conceded this when he repeatedly insisted that we do not know for sure what moves us to act even when we take ourselves to be acting for the best moral reasons. It is clear, however, that Kant meant to deny the Humean thesis that all motivation must stem from sentiments. Insofar as we take ourselves to be moral agents, Kant argues, we must conceive of ourselves as *capable* of being moved by practical reason alone. Sometimes we may be moved by mere sentiment when we think we are guided by reason alone, but we must suppose that we can do what reason requires even if we lack any feeling prompting us to do so. Here Kant goes beyond claims we have explicitly discussed previously, and Kant's view is widely disputed.

There is, however, a way of understanding Kant's point that is less radical than what is usually attributed to him. Kant denies that all action must be motivated by sentiment, feeling, inclination, or sensuous desire, but these terms can be interpreted broadly or narrowly. Similarly, when Kant insists that we can act from reason alone, we can think of "reason" in more or less radical metaphysical ways. If we interpret desires and sentiments narrowly as felt internal pushes and pulls, then Kant's denial that these must be present as motivating causes of all action is more plausible. If, however, we interpret "desire" broadly as just a given disposition to act, then Kant does not deny that we "desire" to follow moral principles. In fact he insists that all moral agents have, inescapably, a predisposition to morality, even though he attributes it to our rational nature rather than our sensuous nature. Again, if "reason" is given a narrow Humean interpretation, it cannot

motivate any act because it is merely an "inert" power to discover natural facts and relations of ideas. But Kant agrees with Hume that reason, so construed (as "theoretical reason"), is not by itself a source of motivation. To have practical reason, according to Kant, is (among other things) to be disposed to acknowledge certain procedural norms for choice, and so in the broad sense it is a kind of "desire" that can figure in practical explanations of why agents choose to act as they do. Humeans question whether these normative commitments are special in ways that warrant attributing them to our nature as *rational*, as opposed to sensuous, beings. Kant, and followers, thinks that there are good reasons for the attribution. This is a dispute that needs more work on both sides; but it is rarely discussed in a fruitful way. This, I think, is largely because Kant's normative position tends to be conflated with his widely rejected appeal to the distinction between noumena and phenomena, to which I turn next.

Fourth, the core ideas of autonomy suggested here also fall short of the most controversial ideas that Kant introduces when he tries to reconcile his ethics with the conclusions he reached in his *Critique of Pure Reason* (Kant 1965). In the third section of the *Groundwork*, and other writings, Kant argues that to attribute to moral agents the sort of freedom of will that morality requires, we must think of them as belonging to an "intelligible world" as well as the "sensible world." The idea of responsible choice employed in practical discussions cannot be reduced to or fully explained by empirical phenomena: a fact that is marked by saying that wills are *noumenal*, in contrast with what is known through experience (the *phenomenal*). Autonomous wills cannot be known as substances in space and time, subject to empirical causal laws. We can "think" but not "comprehend" their existence as "causes" of a nonempirical kind. These are features of Kant's thought that have led many to reject his ethical theory altogether. It is significant, however, that Kant does not start with them as the elements from which to build his ethical theory, even though the views were largely reflected in his earlier *Critique of Pure Reason*. Rather Kant argues first from (supposedly) common moral thought to general normative principles, and only then develops the extreme metaphysical picture (or nonpicture) to square his ethics with the rest of his philosophy. Less radical contemporary interpretations of this aspect of Kant's thought regard it as only an attempt to distinguish two perspectives on human action, the *theoretical/empirical perspective* appropriate to natural science and the *practical/evaluative perspective* when we think about reasons for acting, obligation, and responsibility. This interpretative strategy is to admit that the practical perspective is committed to irreducibly normative ideas, but deny that it is inseparably committed to a faith in mysterious entities outside of space and time. It is not supposed to be a denial of the conclusions of science but another way of thinking and talking about the same human conduct that psychologists study from the empirical perspective. Even this two-perspectives approach is, of course, unconvincing to many critics, and obviously much depends on how in particular the less radical account of the practical conception is spelled out.

Notes

1 Note, for example, that although Kant is committed to the possibility of noumenal "causation" in the *Critique of Pure Reason*, his argument for beginning ethics with an a priori investigation precedes his conclusion that our conception of morality requires us to think of moral agents from a nonempirical standpoint (Kant 1964: 74–81 [406–14] and 118–23 [450–5]).

2 For example, if it is a categorical imperative not to give false witness, then the (moral) reasons not to give false witness override or defeat the consideration that you might make some money by doing so. In other words, all things considered, you should not bear false witness. The question naturally arises, what should one do if two different categorical imperatives conflict? Kant's response was that this is a conceptual impossibility. There cannot be genuine conflicts of duty, only competing grounds or considerations relevant to determining what one's duty is. Thus, if two alleged categorical imperatives give contradictory directions, then we must regard one of them as mistaken, or only valid in a more qualified form. I discuss this problem in more detail in Hill (1996: 167–98). For a somewhat different view, see Donagan (1993: 7–21).

3 I use capital letters to indicate the basic principle, the Categorical Imperative, and small case letters for the derivative principles, categorical imperatives.

4 Aristotle and most other ancient moral philosophers, I think, do not accept that we have reason to be moral only as a means to some desired end independent of it. The Aristotelian view, for example, is apparently that virtue is a constituent part of "happiness," not a mere means to it. We cannot say, either, that he sees moral requirements as "independent" of what promotes our happiness, but it is important that "happiness" for Aristotle is not merely a subjective state or merely an end that we inevitably desire (Annas 1993: 364–84; Hill 1999).

5 In later writings, trying to reconcile his moral philosophy with some minimal religious beliefs, Kant says that, once we determine through reason what our duties are, we can and should think of them as if commands of God (exemplifying pure practical reason). But this does not alter the main point. Duties are not derived from personal orders, should not be followed from fear of punishment or hope of reward, and are binding only because of rational requirements.

6 "Holy will" is Kant's term for the will of any being that is conceived (as God often is) as necessarily willing what is rational, without temptations or the possibility of willing in an irrational way (Kant 1964: 81 [414]). Such a will is perfectly guided by rational principles (regarding what is good) but these principles do not impose imperatives or duties on a holy will. Such a will would be a member of the "kingdom of ends" as a "completely independent being," one whose will (along with the rational will of all members) legislates the moral laws but without being "subject" to the laws as authoritative constraints (Kant 1964: 100–1 [433–4]).

7 Kant thought that acts that express a moral attitude (e.g., a commendable regard for persons as ends) are not acts "from inclinations," such as "pathological love" (i.e., a feeling distinct from commitments of will made for good reasons). So in the case imagined here I am supposing that the love is not so blind and detached from your general commitments and moral attitudes. It is also not a driving force, as conceived on a mechanical model, although it alerts you to concrete needs and you may act with love

(lovingly). My claim now is not that Kant's statements about acts "from duty" are compatible with his acknowledging that our imagined case is a "morally worthy" act, but only that it is a case of acting for the reasons behind recognizing the aid as "duty" and so should have been counted as morally worthy.

References

Allison, H. (1991) *Kant's Theory of Freedom*, Cambridge: Cambridge University Press.

Annas, J. (1993) *Morality of Happiness*, Oxford: Oxford University Press.

Donagan, A. (1993) "Moral Dilemmas, Genuine and Spurious: A Comparative Anatomy," *Ethics* 104: 7–21.

Falk, D. (1986) "Guiding and Goading," in *Ought, Reasons, and Morality*, Ithaca, NY: Cornell University Press, pp. 42–66.

Hill, T. (1992) "The Kantian Conception of Autonomy," in *Dignity and Practical Reason in Kant's Moral Theory*, Ithaca, NY: Cornell University Press, pp. 76–96.

Hill, T. (1996) "Moral Dilemmas, Gaps, and Residues," in *Moral Dilemmas and Moral Theory*, ed. H.E. Mason, New York and Oxford: Oxford University Press, pp. 167–98.

Hill, T. (1999) "Happiness and Human Flourishing in Kant's Ethics," *Social Philosophy and Policy* 16: 143–75.

Hill, T. (2006) "Kantian Normative Ethics," in *The Oxford Handbook of Ethical Theory*, ed. David Copp, Oxford University Press, pp. 480–514.

Kant, I. (1964) *Groundwork of the Metaphysic of Morals*, trans. and ed. H.J. Paton, New York: Harper & Row.

Kant, I. (1965) *Critique of Pure Reason*, trans. Norman Kemp Smith, New York: St. Martin's Press.

Kant, I. (1993) "On a Supposed Right to Lie because of Philanthropic Concerns," in *Grounding of the Metaphysics of Morals*, trans. James Ellington, 3rd edn, Indianapolis: Hackett Publishing Co., pp. 63–7.

Kant, I. (1996) *The Metaphysics of Morals*, trans. Mary J. Gregor, Cambridge: Cambridge University Press.

Kant, I. (1997a) *Anthropology from a Pragmatic Point of View*, trans. Mary J. Gregor, The Hague: Martinus Nijhoff.

Kant, I. (1997b) *Critique of Practical Reason*, trans. Mary J. Gregor, Cambridge: Cambridge University Press.

Schneewind, J.B. (1990) *Moral Philosophy from Montaigne to Kant*, vol. 2, Oxford: Oxford University Press.

Chapter 15

Contractarianism

Geoffrey Sayre-McCord

Introduction

As a general approach to moral and political thought, contractarianism has had a long and distinguished history – its roots are easily traced as far back as Plato's *Republic*, where Glaucon advanced it as a view of justice, and its influential representatives include Grotius, Pufendorf, Hobbes, Locke, Rousseau, Hume, and Kant. In various ways, to various purposes, and against the background of various assumptions, each of these philosophers offered contractarian arguments for the views they defended. What binds the tradition together, in the face of this variety, is the conviction that moral norms or political institutions find legitimacy, when they do, in their ability to secure (under the appropriate conditions) the agreement of those to whom they apply.

As long as the tradition is, it seemed, until recently, to have quietly faded into the past. Several things were responsible for contractarianism's decline. The first was the unsettling realization that even evidently legitimate governments had never actually secured the consent of those governed, which pushed contractarianism to an appeal to the hypothetical consent of hypothetical people under hypothetical circumstances. The second was the rise of utilitarianism and, later, Marxism as substantive alternatives that advanced their own positive views of political legitimacy along with their own criticisms of contractarianism. The third was the influence of positivist criticisms of moral and political theory that seemed to undermine all attempts to develop a rationally defensible normative theory.

Recently, however, contractarianism has enjoyed a dramatic resurgence in popularity. This striking renewal of interest is due not only to the eventual rejection of positivism but, in part, to developments in formal decision and game theory

The Blackwell Guide to Ethical Theory, Second Edition. Edited by Hugh LaFollette and Ingmar Persson.

(developments that promise a clarity and rigor of formulation all too rare in moral theory), in part, to an increasing dissatisfaction with traditional arguments for utilitarianism and its competitors, and, in part, to the sense that individuals deserve a preeminent place in any plausible account of moral and political obligation.

The variety of views that count themselves contractarian is daunting. Some hark back to Hobbes (and his emphasis on a concern with one's own advantage), others to Kant (and his emphasis on a respect for others as ends-in-themselves). Some restrict themselves to subjective theories of value and maximizing conceptions of rationality; others embrace objective theories of value and more capacious accounts of practical reason. Some give a prominent role to game theory and principles of bargaining; others emphasize consensus and conciliation. And some defend familiar forms of utilitarianism (albeit with new foundations), while others defend theories that give a prominent role to rights.

In what follows I hope both to place the contemporary work into historical perspective and to set out some distinctions and contrasts that might help organize, and explain the shape of, contemporary work. However, the historical perspective I offer is not scrupulously historical. I smooth over a good deal of the twists and turns that due care to the historical record would reveal, and I leave out of the account almost completely the various social and political forces that induced not just the twists and turns but the main line of development I identify. For the sake of making clear the positions that emerged as contractarianism developed, and the philosophical considerations that recommended them, the historical picture offered here is more than a little contrived. While the views described emerged (for the most part) in the order suggested and (I believe) for the reasons offered, the succession of views was not nearly so clean as my description implies. Indeed, because each view won devotees long after others saw reasons for change, a number of the positions characterized here as long superseded kept a grip on life even as their offspring thrived and attempted patricide. So, while I describe one view as giving way to another in the face of its perceived weaknesses, I force on the history of contractarian thought an air of inevitable development that is only partially borne out by what actually happened. Still, I hope, this bit of sanitized analytical history will provide a way of thinking about contractarianism that sheds some light on why it developed as it did.

The Background

Contractarianism came into its own in the seventeenth and eighteenth centuries, primarily as a political theory. It developed directly as a response to concerns about the legitimacy of government and the grounds of political obligation. As faith in the divine right of kings evaporated, and the assurance that some were by nature born to rule waned, people came to see political authority as a human creation. The question arose: what could possibly justify the state and explain our obligation

to it? "Man was born free; and everywhere he is in chains," Rousseau (1978: 46) famously noted, "What can make [this change] legitimate?" Social contract theories offered an appealing answer that traced the grounds of political legitimacy and obligation not to God or to nature, but to the wills of the people who were affected. In the process, it promised to articulate the origins of, obligations to, and limits on, legitimate government.

Viewing government either as a conventionally established arrangement among people, or between a people and their sovereign, the social contract was called on in three capacities. First, at least initially, it was offered as an explanation of how governments actually arose. As the explanation would have it, governments emerged in a prepolitical context – the "state of nature" – as solutions to the problems that inevitably arise among people in the absence of a state. Second, the social contract was offered as an account of why people have an obligation of allegiance to their government. The idea was that people have such an obligation because either they have consented, or they have good reason to consent, to the government's authority (as a way to avoid the hardships threatened by the "state of nature"). And third, the social contract was advanced as a justification for limiting the powers of government. The suggestion was that a government's powers are properly limited to those necessary for solving the problems that give rise to government in the first place, since more extensive powers would go beyond what people would have reason to consent to were they in a "state of nature."

Contractarianism provided, as a result, a normative framework that might be used to defend or to attack the legitimacy of particular governments. As it happened, contractarian arguments were in fact relied upon both to foment and to resist revolutionary pressures. Thus Rousseau's *The Social Contract* came to the fore as a natural source for the condemnation of virtually all existing governments while Hobbes' *Leviathan* was, in its implications and intention, conservative in the extreme. When the point was to attack the legitimacy of some particular government, contractarians would argue that the relevant people had no reason to recognize its authority because the life they could realistically expect to face without it would be better. When the point was to defend the legitimacy of some government, contractarians would argue that the relevant people had reason to recognize the authority of some government because the life they could realistically expect to face without the government would be palpably worse.

Originally, and especially while natural law theory still held sway, the contractarian arguments played out against two substantive assumptions: that real consent had (sometimes) actually been given and that people had a moral obligation to keep the agreements they had made. The first assumption legitimated the sovereign's power, since the relevant people were supposed to have given their permission for its exercise. The second established the political obligations of those who were party to the agreement, since they were seen as having undertaken, by way of the agreement, certain specific moral obligations.

Just what constituted giving consent, however, became a ticklish and difficult issue for those who thought it required. Instances of explicit consent seemed rarely

on offer. So tacit consent, given (say) by the acceptance of benefits or participation in certain conventional practices, emerged as the only sort of consent that might actually have been given on a regular basis. Yet tacit consent, if characterized in a way that allowed it to be often enough secured, was so easily given (especially under conditions where no viable options existed) as to be of little use in justifying political authority and obligation. To the extent those too poor to move accepted the benefits of the state, for instance, their tacit consent seemed not so much a reflection of their wills as the inevitable upshot of their situation. Virtually unavoidable, tacit consent seemed insufficient grounds for distinguishing legitimate from illegitimate government.

Questions arose as well about the need for real consent, explicit or otherwise. For it appeared, on the one hand, that particular legitimate governments had never received consent of any kind, and, on the other hand, that whatever might justify a person's moral obligation to keep agreements would serve as well to justify a political obligation (even in the absence of actual agreement) (Hume 1985). Whatever did legitimize government eventually seemed not to depend on the government's consensual pedigree.

Thus actual consent looked to be neither necessary nor sufficient for legitimate government and political obligation. Nonetheless, the idea that people would, with good reason, willingly give their consent, seemed to offer a compelling endorsement of whatever they would agree to – even if they had not actually given their consent. Conversely, the idea that people would, with good reason, not willingly offer their consent, seemed a powerful condemnation – even if their consent had in fact been given either for bad reasons or unwillingly.

As a result, appeals to actual consent, along with the fanciful histories that accompanied early attempts to identify when it had been given, were replaced by appeals to hypothetical consent – appeals to what people would agree to, if only they were rational, and not to what they had actually agreed to. Yet this raised difficulties of its own, for while real agreement might establish real obligations, hypothetical consent, which was no consent at all, apparently established no obligations whatsoever. Still, that people would have given consent, if they were given the chance and were rational, does seem to establish something: that they had reason to support what they would have consented to. So, leaving behind the suggestion that consent (of any sort) was the source of the obligation, contractarians shifted to the claim that what mattered was that people had reason to give their consent if it were needed either to establish or to maintain the government in question. The reasons for giving consent, not the consent itself, were taken as establishing obligation. The authority of some government, in turn, was seen to depend on people having reason to recognize it as authoritative, not on their actually having given that recognition (via consent or contract, or indeed in any other way).

The switch to hypothetical consent (which plays a role in Hobbes and is clear in both Rousseau and Kant) allowed contractarians to avoid explicit consent's implausible histories and tacit consent's excessively lenient account of

commitment, as well as their shared reliance on the assumption that people have a moral obligation to keep their agreements. Appeals to hypothetical consent emphasized instead the reasons agents had for reaching agreement in the first place. And they allowed the theories to rely not on what people might actually have done (perhaps for bad reasons or unwillingly) but on what they had good reason to do.

The reasons people were seen as having were not, importantly, provided by some independently specifiable theory of what constituted legitimate government. The argument was not that people had reason to give their consent because the government was legitimate, but that it was legitimate because they had reason to give their consent. The reasons people (supposedly) had, needed to be found, therefore, in considerations that did not presuppose an account of legitimacy. The considerations offered were, in a straightforward sense, standardly practical, in that they emphasized either the advantages to each that would come from the state, or the ways in which the state spoke to their independent concerns.

Of course the details of the theory shifted substantially as different accounts of the prepolitical state of nature were offered, since those accounts influenced significantly what people might reasonably have agreed to and so what powers they might have recognized as necessary or desirable. Hobbes, for instance, saw the state of nature as so threatening that it called for an absolute sovereign limited only by an inability to demand that citizens willingly submit to death. Whereas Locke, convinced that a prepolitical state of nature would be a tolerably harmonious community suffering primarily from the undesirable effects of people trying to enforce individually what they each regard as their rights, saw the government's legitimate role as really quite limited.

In any case, and whatever the description of the prepolitical state of nature, contractarians offered the description as a realistic characterization of how things would actually be without government. The contractarian approach asked people to consider seriously what life would really be like without government. In an age of revolutions, this was not a call to imaginative flights of fancy; it was the pressing of a seemingly realistic possibility. To the extent the prospect of life without government was palpably worse than life with government, people could see themselves as clearly having reason to consent to some government or other, rather than face the alternative.

Crucial to these contractarian arguments, of course, was their effectiveness in advancing a credible view of how life would be in the absence of government. Those who saw themselves as facing the prospect described saw as well that they had reason to give their consent, if it were required, to some political authority. Against the background of a compelling description of state of nature, the contractarian argument had tremendous appeal.

As long as the state of nature represents a genuine threat, and as long as staying out of that situation requires mutual cooperation in support of government, the real people facing the threat will each find they have reason to recognize the authority of the state. It will not be actual consent that carries the burden so

much as there being compelling grounds for giving it if asked – grounds provided by the recognition that without the government life would be much worse. So while the proof of acceptability was originally thought to be found in actual acceptance, via real consent, that proof came to seem securable, and sufficient, in the absence of consent.

To admit that having some particular government is better than having none, however, is not yet to hold that any government whatsoever is acceptable. Some governments might still be so terrible that people would reasonably prefer no government. Moreover, those governments that do improve on the state of nature will not do so in the same way, and if people faced a choice they might well prefer one of these to the others. No doubt, for any given government, some people falling under it would prefer a different government even if not the state of nature. Yet the question people faced was not just which government (or form of government) would they prefer to a "state of nature," nor was it which would they most prefer given their particular convictions, tastes, skills, weakness, and so on. It was, instead, which government (or form of government) would they prefer from among those that might also secure the support and agreement of the others who are to establish and maintain it. The question became what form of government would properly (and plausibly) secure the agreement of each in light of the fact that no acceptable government could be established without the consent of all (or at least most). Indeed, at the core of contractarianism is the insistence that the arrangements must prove acceptable to all who would fall under them.

To this point, the argument for being willing to give consent was not merely addressed to real people; it was put to them in terms they were supposed to see as accurately characterizing a choice they might in fact have. It was hypothetical consent theory in the face of a possibly real choice. And the reasons people were seen as having were reasons they supposedly actually had in their circumstances.

Unfortunately, while people might willingly consent to (maybe almost any) government rather than face the state of nature, that willingness could well reflect aspects of their actual situation that are morally suspect. That a person has reason to give consent to enslavement rather than face the (perhaps quite real) prospect of painful death at the hands of the would-be master does nothing to establish the legitimacy of the enslavement. Real reasons, under coercive circumstances, may legitimize giving consent, but they will not legitimize others acting on that consent. Thus, in order to justify the authority of some government, hypothetical consent needed to emerge in situations reasonably viewed as morally untainted.

The pressure to purify the circumstances of agreement naturally led the hypothetical consent contractarians to advance as well idealized circumstances for that consent. Yet some resisted an appeal to idealized circumstances, for one of two reasons. On the one hand, some thought the actual circumstances, and the "state of nature," are appropriately untainted even if life is less good than it might be. On the other hand, some thought that whatever reasons a person might have for making an agreement under idealized conditions, those reasons would be irrelevant to real agents unless those reasons were likewise reasons the agents would

have under suitably realistic conditions – in which case there was no point in appealing to the idealization. Still, most contractarians were moved to hold that the relevant choice situation was one made under idealized conditions, in which those who gave consent were free from coercion.

Similar considerations worked to recommend too that the people whose consent mattered were not people as they actually are – sometimes irrational and often ignorant – but those people as they would be were they (for instance) perfectly rational and appropriately informed. That a person, when irrational or ignorant, would (or even does) give consent, under even noncoercive circumstances, to some arrangement, does nothing to establish the person has reason to give that consent. Consent under noncoercive circumstances may legitimize others acting on that consent, but it will not legitimize thinking the person has reason to consent.

All the while recognizing that the question is whether real people have reason to endorse the government whose legitimacy is at issue, contractarians began to distinguish these real people, and their actual circumstances, from the (suitably idealized) people who are supposed to reach agreement and the (appropriately idealized) circumstances under which their agreement was supposed to be secured. That suitably idealized people would, under appropriately fair circumstances, willingly agree to some (form of) government, came to seem the plausible standard of legitimacy. What grounds could there possibly be, one is inclined to ask, for objecting to such a government?

A natural worry, however, is that the rhetorical force of this question was bought at the expense of it being vacuous. For it looks as if all the interesting justificatory work would be done in specifying who might count as suitable parties to the agreement, and what would count as appropriately fair circumstances. The contractarian test came to seem empty without the addition of some noncontractarian theory that would identify not only who counts as suitably rational, and what circumstances are appropriately noncoercive, but also what reasons there are for consenting under those circumstances. Such a theory would presumably support its own substantive account of political legitimacy independently of having to appeal to who might consent to what. Whereas, initially, contractarianism appealed to real people in their actual circumstances facing a real choice, it now was so removed from the real world, and so normatively laden in its assumptions, that the contractarian framework seemed at best a useful heuristic for discovering some independently specifiable criterion that must be defended on some other, noncontractarian, grounds.

A number of theories emerged as candidates for the role. The most influential, early on, was utilitarianism, which held, in effect, that what rational people would agree to (under actual as well as idealized circumstances) is precisely whichever government would maximize overall welfare. The legitimacy of a government turned, utilitarians argued, on how well the government advanced the interests of everyone concerned. And the reason any particular person had for recognizing the legitimacy of the government – and so for consenting to its authority – was traceable not (for instance) to how that person would fare, but to how people in general

would. Of course, this theory did not require the contractarian framework for its articulation or deployment. Yet if one asked whether, on the utilitarian view, all rational people could, under appropriately fair conditions, willingly give their consent to the (form of) government it endorsed, the answer was an easy "yes."

Other theories, such as natural rights theories and (on some interpretations) Marxism, also stepped into the breach and offered accounts of what might constitute fair conditions of, and good reasons for, agreement. Each of them, though, was in a position to sidestep an appeal to the contractarian framework even as it had the resources to say that all (really) rational people would, in the appropriate circumstances, willingly agree to the (form of) government the theory legitimized. To the extent these theories did not co-opt the rhetorical force of the contractarian framework, they were used instead to undermine it, on the grounds (for instance) that the framework illegitimately valorized the individual or ignored the value of community or substituted market relations for moral ones (see Pateman 1988; Sandel 1982).

Once contractarianism stepped away from reliance on the real consent of real people, and once it moved on to embrace as relevant only the hypothetical consent of idealized people in idealized circumstances, it not only invited noncontractarian additions, it seemed to need them. And once the additions were at hand, they did not just add to contractarianism, they displaced it.

Recent Contractarianism

That is pretty much where things stood until the middle of the twentieth century; although, in the first half of this century, things got even worse for contractarianism, thanks to the influence of logical positivism. For according to the positivists, grand attempts at moral theory and political justification are, despite pretensions to the contrary, actually merely elaborate devices for bringing others onto one's own side, not articulations of independent standards we might discover and defend rationally (Ayer 1936).

The relatively recent revival of contractarianism has depended upon an emerging conviction that, contra positivism, there must be room for reasoned argument about normative matters. But the revival has required two other things as well: first, a growing dissatisfaction with the moral theories that had displaced contractarianism, and, second, the prospect that recognizably contractarian considerations might after all contribute nontrivially to moral theory.

Moreover, just as political contractarianism emerged as a response to the recognition that political legitimacy and obligation could not be traced to God or nature, moral contractarianism's appeal has grown substantially with the sense that moral constraints must in some way be a reflection of human reason or social convention, not of God or (nonhuman) nature. Contractarianism holds out the seductive prospect of a theory that demystifies morality's status and shows it to be

a compelling expression of humanity's nature. For if morality finds its source and authority in our capacity to embrace its demands, then understanding morality will ultimately require appealing to what we would need in any case in order to explain our own capacities and practices. Nothing occult or mysterious or supernatural need be implicated (Mackie 1977; Milo 1995).

The contractarian framework, with its appeal to what people would agree to under appropriate circumstances, has found a natural home in two very different approaches that take their inspiration (though frequently little else) from Kant and Hobbes. The Kantian approach begins with our natural concern with morality and uses the contractarian framework to specify and draw out the implications of that concern. Contractarianism, in this case, is advanced as a way to articulate the content of morality. The Hobbesian approach, in contrast, acknowledges our concern with morality but sees that concern itself as properly called into question and uses the contractarian framework to explain why, and to what extent, we have (nonmoral) reason to embrace morality. Contractarianism, in this case, is advanced as a way to justify a concern for morality's content and demands. On either approach, contractarianism's distinctive commitment to seeing legitimacy as grounded in what people might willingly agree to under the appropriate circumstances finds a central role.

Kantian Contractarianism

The Kantian approach has famously been pursued by John Rawls, who introduces the contractarian framework to articulate morality's impartiality. In the process, he hopes to take "seriously the distinction between persons" in a way that other attempts to capture impartiality do not (1971: 187).

That moral demands are impartial is, of course, acknowledged by virtually all moral theories. Yet the nature of that impartiality and its implications for morality's authority and content are quite controversial.

One familiar way to capture morality's impartiality is to suppose that moral demands flow from a source equally concerned for all who fall under them. Thus many religious views portray morality's demands as reflecting God's equal love for all, while Ideal Observer theories treat the demands as an expression of what an equisympathetic observer would approve of, and utilitarian views see the demands as giving equal weight to the welfare of all. This way of capturing impartiality leads naturally (although not inevitably) to moral principles that are decidedly utilitarian in their implications.

Another way of capturing morality's impartiality, however, is to see its principles as those we would each, individually, choose to govern everyone's behavior if our choice was made in ignorance of how we, as we actually are, might benefit or suffer as a result (Harsanyi 1953; Rawls 1971). With this in mind, Rawls describes the appropriate circumstances of agreement – the relevant "state of nature" – as including a "veil of ignorance" that shields from view all information concerning

the particular talents, tastes, history, and situation of those seeking agreement. Impartiality is achieved by eliminating all the information that would engage partial concern. Collective and partial choice under circumstances of radical ignorance is substituted for individual choice under circumstances of extraordinary impartiality and knowledge. And with the substitution comes the appeal to contractarianism. For now the legitimacy of certain principles, and their standing as distinctively moral, appropriately impartial, principles, turns on whether people would – under the relevant circumstances (in this case, circumstances of ignorance that neutralize partiality) – choose the principles. Hypothetical choice, under hypothetical circumstances, sets the standard for moral legitimacy, on this view, because such choice embodies impartiality.

One apparently significant advantage of the contractarian approach to impartiality is that it need not appeal either to interpersonal utility comparisons or to any general method of balancing the interests or welfare of people. In contrast, when impartiality is embodied in equal love, or sympathy, or concern for other's welfare, an appeal to interpersonal utility comparisons and an overall balance of advantages seems inevitable. Many see this difference as grounds for thinking that the contractarian embodiment of impartiality captures especially well the moral significance of the individual and the idea that one person's loss cannot always be morally compensated by another's gain (Rawls 1971; but see Harsanyi 1953).

Impartiality is only one aspect of morality that invites contractarian elaboration. Rather than starting from the conviction that moral demands are impartial or fair, some versions of contemporary contractarianism focus instead on the idea that moral reasons are public and shared – they provide reasons for all. These approaches shift attention away from conflicting interests that call for impartial arbitration towards a collective concern to accept principles all can embrace as reasonable. Contractarianism's appeal to mutual agreement (under appropriate circumstances) strikes many as doing a uniquely satisfying job of articulating the sense in which morality's demands can lay claim to the allegiance of all. Indeed, by treating moral norms as just those everyone has reason to accept, contractarianism not only articulates the connection between morality and mutual acceptability but takes that connection as definitive. A concern to act morally, on this view, is a concern to act in light of principles that everyone might reasonably embrace. To determine what these principles are we need to ask the distinctively contractarian question: to what could people, under the appropriate circumstances, reasonably agree? This time, though, the appropriate circumstances are conceived not as involving radical ignorance but instead as being occupied by participants who offer considerations for and against various principles in a context where all are supposed to be both reasonable and concerned to settle on principles all the participants can accept. (See Scanlon 1998; Habermas 1990.)

While impartiality and mutual acceptability have both played crucial roles in making contemporary contractarianism attractive, they themselves find support in a third aspect of morality – the evident importance of equal concern and respect. Many hold that the moral significance of individuals is best captured by a view

that treats morality's demands as themselves a reflection of what each person, uncoerced and conceived of as a full participant in the process, could rationally embrace. To treat a people with equal concern and respect, on this view, is to see them, no less than oneself, as having a legitimate say in the principles that should govern your interactions. By governing oneself by principles others could endorse, one thereby gives expression to the equal concern and respect that is distinctive of morality.

As I have suggested, contractarians disagree among themselves as to which, if any, of these various considerations ought to be given primacy. Even among those who agree on that, there is significant disagreement as to how impartiality, or mutual acceptability, or equal concern and respect, might find their best expression. Despite the disagreement, however, there is consensus among those taking this approach that the relevance of the contractarian framework is found in its capacity to articulate crucial and distinctive features of morality. Thought of in these terms, contractarianism addresses those who are already concerned to do as morality demands but are trying to figure out what, precisely, that might be. Contractarianism is offered as a way of specifying those demands and is defended as appropriate by appeal to its capacity to articulate and embody crucial features of morality.

The Kantian approach to contractarianism faces two related problems. One concerns whether, in the end, any real work is being done by the appeal to the agreement of people, properly situated. The more successful an account is in eliminating the influence of individual differences on choice, in the name of impartiality, the less room there seems to be for the idea that the choices of distinct individuals matter to the outcome. When it comes to choices behind a veil of ignorance, for instance, asking what all might agree to under that circumstance appears functionally equivalent to asking what any one person might agree to, since the veil hides from the scene all the features of a person that might distinguish one person from others. Do the notions of collective choice or mutual agreement really have any substantive place in such theories? It is not at all clear. In any case, a second familiar problem emerges when one asks on what basis the people are to reach a collective choice or mutual agreement, supposing they do. Any grounds people who are "properly situated" might have for settling on one choice or agreement rather than others threaten to stand independently of any choice or agreement at all. Even hypothetical agreement among people seems to drop out of the picture. The concern underlying both of these problems is that the contractarian appeal to what people – in the plural – might agree to, under whatever circumstances, seems not to be playing anything other than a heuristic role.

This is a serious concern. If it is not met, the contractarian framework will stand as mere window dressing – a decorative overlay that might have evocative advantages but that contributes not at all to the substance of a theory or the justification of the principles it endorses. The central challenge is to find a role for the distinctively contractarian idea that morality's demands reflect, in some nontrivial way, what people might reasonably agree to under the appropriate circumstances.

Of course, a range of obviously noncontractarian theories can end up saying that the principles they advance might be chosen by reasonable people properly situated. When they do say this, however, the appeal to what the people might agree to (or choose) swings off the side as a fifth wheel rolling along with the theory but driving no part of it. The content of the principles advanced is wholly unaffected by consideration of what people might reasonably agree to – all the influence goes the other direction.

In order to meet the central challenge, a contractarian theory has to show that the appeal to what people might agree to is sensitive in some way either (1) to the variety of people participating in the choice, or (2) to the variety of people to be governed by what is chosen, or (3) to the variety of people being addressed by the argument. Only then will the idea that morality turns on what distinct individuals might agree to have a significant role. If, alternatively, the key choice or agreement in play might as well be made by and for a single individual, talk of what people (as opposed to a person) might agree to will have no substantive impact on the nature of the principles that are supported by the argument, and the contractarian framework will be making no distinctive contribution to the theory. As it happens, all three of these options have been explored, exploited, and defended.

Thus some contractarians argue that the appropriate circumstances of choice leave intact, in the way a veil of ignorance might not, the key fact that those who are seeking to reach agreement differ from one another in ways that influence which principles might be genuine candidates for mutual agreement. This sort of argument usually characterizes the relevant agreement as being a result of bargaining or some kind of balanced accommodation, the particular content of which reflects differences among the participants. By leaving intact individual differences and allowing those differences to have some impact on the nature of the principles that are supported by the argument, such views make genuine room for the contractarian thought that the legitimacy of certain principles depends nontrivially on what people collectively might agree to.

Other contractarians argue that the outcome of the choice in question is shaped substantially by the fact that what is being chosen (a set of principles to govern interactions among individuals, or a set of basic institutions to structure society, or whatever) is chosen for a potentially diverse group of people who differ in talents, values, personality, and so on. This sort of argument turns our attention to the ways in which the choice problem is shaped by the prospect of the results applying to different people. Even if, in the appropriate circumstances, one chooser is as good as another and more than one is no real addition (as might be the case behind a veil of ignorance), it may be that what such a chooser would select is influenced by the fact that those for whom she is choosing are different in ways that need to be accommodated, from the start, by the choice she makes.

Still other contractarians argue that the whole choice situation – the circumstances under which it is to be made and the nature of those who are to make it – is answerable, in a nontrivial way, to the fact that the overall contractarian

argument is being offered not to a single person but to people insofar as they see themselves as together trying to settle on acceptable principles for interacting with each other. This sort of argument focuses on the situation of the actual people to whom the arguments are addressed and maintains that their differences have an impact on just how the choice situation is to be described. A crucial feature of our actual situation, it seems, is that we can expect reasonable people to disagree fundamentally about central philosophical and moral issues in ways that mean there is no real prospect of reaching an across the board consensus. Nonetheless, there may be room for all reasonable people to agree (for their different reasons) that there are reasons to regulate our interactions by norms that are mutually acceptable by all who are reasonable and we may see the original contract situation as articulating the common ground shared by all those who are reasonable (Rawls 1993).

All three lines of argument have more than a little plausibility. At the very least, they suggest there might be room to defend the view that the contractarian framework, in some guise or other, can play more than a heuristic role in a theory. Still, those who offer some version of contractarianism as the best articulation of moral concerns we are assumed to share face two additional worries.

The first is that the variety of contractarian theories itself testifies to the fact that people's prior understandings of morality differ significantly, even when it comes to thinking through what impartiality, say, consists in. And this raises a problem since, in the face of a range of different contractarian theories, the question naturally arises: which, if any, of these articulations of our prior concern captures accurately the object of that concern? Asking what people, appropriately situated, might agree to seems to provide no purchase whatsoever on that question, since all the different versions of contractarianism will travel with their own preferred description of the circumstances of choice and of the grounds on which the people so situated will reach agreement. Whatever counts as good grounds for settling on one (contractarian) characterization of our moral concern rather than another, it seems it will not be grounds that turn in any interesting sense on what people would agree to, if they were properly situated. The fundamental argument for one view rather than another looks as if it will have to be decidedly noncontractarian.

The second worry arises even if a particular characterization of the contractarian framework emerges as successfully capturing our moral concerns. It centers on the question: what reason is there to embrace that moral concern? Even those who are concerned to act as morality requires might, on reflection, wonder whether they have any good reason to retain or act on that concern, especially in situations where morality quite clearly requires sacrifice. Why not think of the concern as merely a reflection of socialization that one would do better to be without? If contractarianism is offered solely as a way to articulate a concern for morality that we are assumed to share, it will in effect ignore the issue. But many think this is an issue that should not be put to one side casually, not least of all because so often people's actual concerns reflect ignorance, superstition, and prejudice.

Morality, of course, presents itself as legitimately commanding allegiance and sacrifice. But do we really have reason to offer the allegiance and make the sacrifices, when called for?

The Hobbesian approach to contractarianism takes this challenge seriously and sees the contractarian framework as offering a uniquely compelling answer to it. Before turning to this approach, though, I should mention one tempting answer that is available to the Kantian contractarian. As this answer would have it, those who acknowledge that some course of action is morally right or required, but wonder whether they have reason to act accordingly, are failing to appreciate something that follows directly from what they have acknowledged: that they have (moral) reason to act as required. Acknowledging a moral demand, it seems reasonable to think, carries in its wake recognition of a moral reason to act accordingly. But this observation just pushes the problem back a step, since now the question is whether one has any good reason to give weight to moral reasons in one's deliberations. One might insist at this point that moral reasons are necessarily weighty so that once we admit there are moral reasons there is no good sense to be given to wondering about the weight of those reasons. Yet it is not hard to sympathize with those who would feel cheated by such an answer – cheated of a non-question-begging defense of the importance of morality.

Hobbesian Contractarianism

The Hobbesian approach to contractarianism offers such a defense. Those who take this approach argue that we have nonmoral reasons to embrace morality. The distinctively contractarian element in this approach comes with explaining the way in which the reasons we each have for embracing morality are reasons that reflect the interdependence of our interests and the opportunities we have for mutual benefit. Speaking in broad terms, the Hobbesian approach views morality as constituted by a set of principles the adoption of which is advantageous for everyone in a way that means each person would have (nonmoral) reason to adopt the principles as long as others did as well. And it sees the legitimacy of morality's demands as turning on our having (nonmoral) reason to support them.

One of the earliest versions of contractarianism, advanced in Plato's *Republic*, contained the core elements of the Hobbesian strand of contemporary moral contractarianism. Put in Glaucon's mouth, this version of contractarianism is pleasingly direct. According to the view he sets out, the rules of justice are conventional and represent a compromise. On the one hand, people would prefer to have their wills unchecked by others. On the other hand, they would prefer not to suffer the unchecked wills of others. Recognizing that they cannot enjoy the first without suffering the second, and rightly fearing the second, they band together to establish and enforce mutually agreeable limits on each other's wills. These limits, Glaucon suggests, simply are the rules of justice. Thus, as Glaucon tells the story, the constraints justice imposes are a reflection of convention, yet the reflection of

a convention we each have (nonmoral) reason to encourage and embrace (given the human condition). The convention that constitutes morality, while a compromise of sorts, is nonetheless a reasonable one that redounds to our mutual advantage. (Gauthier 1986, Buchanan 1975, and Harman 1978, all offer contemporary defenses of this sort of view.)

Game theory provides resources for representing perspicuously the underlying structure of social interactions that give point, in the way Glaucon suggests, to moral principles. In the process it has made possible a sophisticated investigation of the various different ways in which the reasons any particular person might have to act in one way or another depend upon what others have reason to do. Perhaps most influential on this front has been the Prisoners Dilemma, which models a situation where the options and available benefits are such that, if each person directly maximizes her expected utility, they will together predictably end up worse off than they would have been had they cooperatively forgone immediately available benefits.[1] But various other notions from game theory and economics have played crucial roles in recent discussions of contractarianism. Especially important on this front have been developments concerning the understanding of free riders (who enjoy a benefit thanks to the efforts of others without themselves participating in producing that benefit), externalities (which are costs imposed by decisions that are shifted to those who have had no say in their production), and assurance problems (where a potentially available benefit for all will be beyond reach unless all have assurance that others will do their parts). In each of these cases, it looks as if a successfully established and internalized set of principles requiring certain kinds of action, demanding the consideration of others, and underwriting confidence that others will act in concert, would alleviate problems we would all otherwise face. Reciprocal constraints, intelligently selected, lead to mutual advantage.

The hope held out by Hobbesian contractarianism is that, at least to some extent, moral principles might ultimately be justified by showing the extent to which we all benefit from living in a community of people who constrain their pursuit of interest by those principles. At the same time, though, the hope must be balanced by the recognition that in many ways the Hobbesian approach will likely support principles that match common sense at best only imperfectly.

On the Hobbesian view, for instance, the advantages we each enjoy from morality come primarily from others embracing moral principles and secondarily from our avoiding the burdens we would suffer were others to punish us for violating those principles. In particular situations, the balance of advantages may fall in favor of violating particular principles, especially if one can do so undetected (and so unpunished) by others. As a result, even where there might be mutual advantage in establishing recognizably moral principles to govern our interactions, there may in certain cases be no advantage from – and so no reason, on this view, for – compliance. And even when there is an advantage to be gained, the motive for so complying appears to be distinctly nonmoral. Thus at most the Hobbesian approach seems to underwrite acting morally for nonmoral, indeed apparently

selfish, reasons. To the extent a full justification of being moral involved justifying doing as morality requires *because* morality requires it, the Hobbesian might seem incapable of providing the justification (but see Gauthier 1986 and Sayre-McCord 1989).

Moreover, on this view, the principles that would be mutually advantageous overlap only contingently, and then pretty clearly only partially, with those we currently recognize as moral principles. After all, to the extent the principles we have reason to embrace turn on what others too have (nonmoral) reason to embrace, the principles will almost surely reflect the differential power, wealth, and general situation of those party to the arrangement. Similarly, when it comes to those whose protection brings no advantage to others (e.g., the weak and infirm), the principles, the adoption of which would bring mutual benefit, would presumably not offer them protection, since such protection would bring no advantage to others. In both cases, the resulting – mutually advantageous – principles will presumably differ from those recommended by commonsense morality.

The tension between commonsense morality and the principles that would be recommended by Hobbesian contractarian is due in no small part to holding that principles are legitimate only if they can be shown to be advantageous to real people in their actual circumstances. For, almost inevitably, morally suspect differences among people will then influence the content of the principles that will qualify as legitimate (because genuinely advantageous). Yet the more one corrects for these morally suspect differences by focusing not on actual advantages people might expect but on the advantages that would be secured under hypothetical circumstances, the less one can claim real people have (nonmoral) reason to care. If I have been born to comfort and wealth, or have secured such a life through force or fraud or cunning, any subsequent agreements I might make would no doubt be distorted by my initial advantages. Of course someone might, on grounds of fairness, say, insist on disallowing the influence of these advantages, but then the prospect of mutual advantage plummets as the real benefits to me disappear. The moral appeal of the resulting principles seems to be inversely proportional to their claim on actually being advantageous to all. Be that as it may, if one takes seriously the idea that people should act as they have reason to, and if one thinks what one has reason to do is whatever is personally advantageous, then a mismatch between advantage and commonsense morality would be all the worse for common sense.

Fortunately, Hobbesian contractarians can and usually do admit that people's interests and preferences may be other-regarding, sympathetically directed, and broadly sensitive in ways that mean a true appraisal of how their interests are intertwined with other's will reveal, after all, an argument from mutual advantage to principles that are recognizably moral. Their appeal to interest and advantage in defending moral principles does not have to be an appeal solely to self-interest and private advantage. By taking honest account of human nature, and the extent to which we can be engaged by the welfare of others (to a greater or

lesser degree, in response to both nature and nurture), it is at least plausible to think real advantage for all may be secured by the adoption of moral principles already securely established in common sense.

No doubt any defense of morality that needs to appeal, in this way, to our fellow feeling leaves moral principles contingent in two ways that may be disturbing. First of all, the content of the principles is, on such a view, contingent upon the existence and shape of our concern for others. Second of all, the force of the argument offered for giving allegiance to the principles will be contingent as well on the actual concerns of those addressed; although, unlike Kantian contractarianism, the Hobbesian variety need not suppose that the people addressed by the argument already possess a distinctively moral concern.

Significantly, the Kantian and Hobbesian approaches may complement rather than compete with each other. For it may well be that the concern we have nonmoral reason to embrace (as the Hobbesian would argue) is a distinctly moral concern, the content of which calls for contractarian elaboration (as the Kantian would maintain).

Humean Contractarianism

There is a third version of contractarianism, inspired by Hume, that takes for granted neither a concern for morality nor any particular account of what people have reason to do or accept (Hume 1978). It sets out to explain generally why evaluative concepts and commitments (concerning not simply morality, or justice, but also concerning what people have reason to do and to agree to) would naturally emerge among beings with our capacities, concerns, strengths and weaknesses. The distinctly contractarian elements in the etiological story revolve around the evaluative concepts and commitments themselves being conventional solutions to problems people would otherwise face. In setting out to explain our evaluative concepts without presupposing others the Humean contractarianism is more ambitious than either the Kantian or the Hobbesian approaches. Yet the ambition is mitigated by the fact that the Humean approach is concerned neither to establish any particular substantive moral view nor to argue that people have reason, of any particular sort, to be moral. Instead, it hopes to account for the evaluative concepts we actually possess (contractarian in content or not, rationally embraced or not) by appealing initially only to nonevaluative features of our situation and the de facto advantages that come with the capacity to think in evaluative terms.

Once evaluative concepts are up and running, and have a life of their own, the Humean – no less than others – will rely on them in justifying or criticizing not only particular actions and institutions but also, in some cases, the conventions that give shape to the concepts themselves. As a result, the Humean contractarian might well end up defending a particular evaluative stance concerning morality and practical reasons more broadly. So she may embrace the Kantian view that we possess a moral concern that is best articulated by appeal to the contractarian

framework, or she may share the Hobbesian view that a proper understanding of what people have reason to do shows that their reasons are essentially bound up with their own advantage. Or she may reject both views. Her commitment is to seeing the evaluative concepts she relies on as being grounded in, and shaped by, a distinctive set of conventions. Just as moves in a game of chess make sense only in the context defined by the rules of the game, so too, the Humean maintains, evaluative judgments make sense only in the context defined by the conventional rules governing the concepts that are deployed in those judgments. Our capacity to think in moral terms and to talk of reasons depends, on this view, upon resources that are available only once certain conventions and practices have been established.

Significantly, Humeanism offers an account of the conventions that give place and point to distinctively evaluative concepts, not (or at least not merely) an account of fellow feeling or altruism or cooperative dispositions. Thus it hopes to explain our capacity to make evaluative judgments (moral and otherwise) and not merely our capacity to get along or respond emotionally. Presumably the conventions that define our evaluative concepts require the presence of various affective reactions and dispositions. But the Humean's focus is on those conventions themselves and the way in which they serve to constitute our evaluative concepts by setting standards for their correct application.

The Humean approach resembles the Hobbesian, in that the introduction of evaluative concepts (and the principles or standards that specify their content) is seen as an advantageous solution to a problem people collectively would otherwise face. Yet there are some crucial differences. In particular, on the Humean view, while the concepts do have this benefit they are not seen as deliberately introduced on the basis of reasons people recognize (since, by hypothesis, there is no substantive concept of reason yet in play and so no sense to be made of people actually recognizing reasons). Thus, in the first instance, the concepts are seen as arising in an explicable way, given the situations in which people would find themselves, but not as solutions to a problem of collective choice that are embraced by people who see them as rational. Of course, once the relevant concepts are in place, it is possible to reflect back on the introduction and evolution of the evaluative concepts. And on reflection, the Humean approach supposes, one will discover there are good grounds for being glad something like the original concepts were introduced and for endorsing what they have become as they have evolved. However, at this point, the grounds for approving of the evaluative concepts we share will go beyond the austere resources of an appeal to self-interest and will implicate substantive considerations of fairness, justice, and value in ways that a Hobbesian excludes from consideration. (See Sayre-McCord 1994.)

Of course, reflecting on the origin and nature of our evaluative principles may well reveal deep problems with our current understanding of our evaluative concepts. But then the grounds for criticism and the justifications offered for altering our understanding of what justice, say, requires, will of necessity invoke evaluative concepts we have and can understand. The original, mutually advantageous,

conventions will be providing, in these cases, both the resources and the reasons for reflectively correcting the conventions as they stand. The process of reflective adaptation that is then in play is a crucial element in making sense of an otherwise puzzling and anyway distinctive feature of evaluative concepts – their essential contestability.

The process of reflective adaptation plays an important role in addressing two worries. The first worry is that Humean conventionalism is committed to an objectionable form of relativism since the concepts that may emerge in one community might well differ substantially from those in another community. The second worry is that by giving a central role to mutual advantage the approach will inevitably underwrite moral principles that are arbitrarily parochial in their focus and implications. After all, the concepts that do emerge in a particular society, it seems, will be shaped by the interests of those in the community without regard to others. Both considerations raise serious worries, of course, but only insofar as the relativism involved is objectionable and the parochialism arbitrary. There is no doubt that some versions of relativism are objectionable and that parochial concerns are often arbitrary. Still, that different communities may develop different evaluative concepts to answer to their particular situations seems not only something that obviously happens, but also unobjectionable (as long as the concepts in questions are unobjectionable). Similarly, that the concepts that develop within a community answer to that community's needs and interests seems not at all arbitrary. Nor does it seem disturbing on other grounds once we notice that the content of the concepts we have an interest in having may well, and in fact do, bring within their scope the interests of others. To the extent there are reasons to expand the scope of our principled concern or adjust our understanding of our commitment's implications, those reasons are articulated using our current concepts. And these are concepts the currency of which finds an explanation in the Humean story of their social role. Our capacity to criticize principles and practices cannot outstrip the conceptual resources we have for identifying and articulating the supposed difficulties. What the Humean view offers is an explanation – a metaphysically and epistemically modest explanation – of those resources. Barring the discovery that our evaluative concepts carry the seeds of their own destruction, Humean contractarianism is well placed to accommodate and even embrace whatever substantive considerations might be mobilized for thinking there is reason to reevaluate our evaluative concepts.

This very capacity to accommodate and adapt to new considerations calls into question the value of Humean contractarianism, to the extent one hopes to use contractarianism to identify and defend some particular (and fixed) set of evaluative principles. It is important to recognize that this approach cannot offer, and does not pretend to offer, such a defense. Instead, the aim of Humean contractarianism is to explain the origin and nature of our evaluative concepts in a way that shows them to find their source in human nature. At the same time, though, the hope is to show that in discovering the origin and nature of evaluative principles we simultaneously show them, thereby, to have a claim on our allegiance. In playing

the role they do in social interaction, in serving as the medium (so to speak) through which people can coordinate actions and recommendations and resolve conflicts, our evaluative concepts at least in part earn their own endorsement.

Conclusion

Whether and how the Humean approach might mesh with the Kantian and Hobbesian approaches to contractarianism is unsettled. Those tempted by Humean contractarianism, myself included, suspect that it can offer a philosophically satisfying account of (1) when the Kantian appeal to what people might find mutually agreeable under fair conditions is relevant to determining moral demands (and when it is not) and (2) why the Hobbesian appeal to our nonmoral interest in resolving conflict and coordinating behavior is relevant to morality's demands (but why it does not ultimately limit their scope).

Whether and how contractarianism of any sort might ultimately be defended is also unsettled. However, those tempted by contractarianism suspect that a proper understanding of morality must see morality as a reflection of what those subject to its demands might reasonably accept.[2]

Notes

1 See Luce and Raiffa (1957) for the classic description of the dilemma. Here is the dilemma. Imagine two prisoners find themselves facing the following offer. If neither confesses to the crime they are being charged with, they will both be convicted of some lesser crime (that carries a penalty, let us say, of a year in prison). If they both confess, then both will be convicted of the more serious crime but will receive some leniency for having confessed (so they will each, say, serve five years). But if one confesses and the other does not, the person who confesses will get off free with no penalty and the other will serve the maximum sentence (of, say, ten years). Assuming the various years in prison represent costs to the individuals in proportion as they add up, the prisoners face a kind of dilemma: each sees that whether the other person confesses or not she does better to confess, since if the other person confesses she can, by confessing herself, spend only five years, rather than ten, in prison and if the other person does not confess she can, by confessing herself, spend no time at all in prison rather than a year. But if each acts according to this reasoning, they will together confess themselves into five years of jail each rather than the one that they would have been sentenced had they both kept quiet. Yet as soon as one has reason to think the other will not confess she finds herself again with compelling reason to confess. . . . The structure of the dilemma remains even if the costs and benefits at stake are radically different and regardless of whether they represent the selfish concern of the criminal simply to stay out of jail or the selfless preoccupation with worries about the welfare of her children. Moreover, assuming the payoffs have the Prisoners' Dilemma structure, prior promises to remain silent leave the dilemma in place. What is needed to solve it is either something that

will change the payoffs so as to eliminate the dilemma or grounds for reasoning in some way that breaks free from simply maximizing expected utility.

2 Thanks are due to Robert Goodin, Philip Pettit, Michael Ridge, and Michael Smith for comments on an earlier draft of this essay.

References

Ayer, A.J. (1936) *Language, Truth and Logic*, London: Gollancz.

Buchanan, James (1975) *The Limits of Liberty*, Chicago: University of Chicago Press.

Gauthier, David (1986) *Morals By Agreement*, Oxford: Clarendon Press.

Habermas, Jürgen (1990) "Discourse Ethics: Notes on a Program of Philosophical Justification," in *Moral Consciousness and Communicative Action*, trans. Christian Lenhardt and Shierry Weber Nicholasen, Cambridge: MIT Press, pp. 43–115.

Harman, Gilbert (1978) "Relativistic Ethics: Morality as Politics," *Midwest Studies in Philosophy* 3:109–21.

Harsanyi, John (1953) "Cardinal Utility in Welfare Economics and the Theory of Risk-Taking," *Journal of Political Economy* 61: 309–21.

Hume, David (1978) *A Treatise of Human Nature*, Oxford: Oxford University Press.

Hume, David (1985) "Of the Original Contract," in *Essays: Moral, Political and Literary*, ed. Eugene Miller, Indianapolis: Liberty Classics, pp. 465–87.

Luce, R.D. and Raiffa, Howard (1957) *Games and Decisions*, New York: John Wiley & Sons, Inc.

Mackie, J.L. (1977) *Ethics: Inventing Right and Wrong*, Harmondsworth, UK: Penguin Books.

Milo, Ronald (1995) "Contractarian Constructivism," *Journal of Philosophy* 92: 181–204.

Pateman, Carole (1988) *The Sexual Contract*, Stanford: Stanford University Press.

Rawls, John (1971) *A Theory of Justice*, Cambridge, MA: Harvard University Press.

Rawls, John (1993) *Political Liberalism*, New York: Columbia University Press.

Rousseau, Jean-Jacque (1978) *On the Social Contract*, trans. Roger and Judith Masters, New York: St. Martin's Press.

Sandel, Michael (1982) *Liberalism and the Limits of Justice*, Cambridge: Cambridge University Press.

Sayre-McCord, Geoffrey (1989) "Deception and Reasons to Be Moral," *American Philosophical Quarterly* 26: 113–22.

Sayre-McCord, Geoffrey (1994) "On Why Hume's 'General Point of View' Isn't Ideal – and Shouldn't Be," *Social Philosophy & Policy* 11: 202–28.

Scanlon, Thomas (1998) *What We Owe to Each Other*, Cambridge, MA: Harvard University Press.

Further Reading

Gauthier, David (1991) "Why Contractarianism?" in *Contractarianism and Rational Choice*, ed. Peter Vallentyne, New York: Cambridge University Press, pp. 15–30.

Gough, J.W. (1957) *The Social Contract*, 2nd edn, Oxford: Clarendon Press.

Harsanyi, John (1976) *Essays on Ethics, Social Behavior, and Scientific Explanation*, Dordrecht: D. Reidel.

Kant, Immanuel (1964) *Groundwork of the Metaphysic of Morals*, trans. H.J. Paton, New York: Harper & Row.

Kant, Immanuel (1970) "On the Common Saying: 'This May be True in Theory, But It Doesn't Apply in Practice'," in *Kant's Political Writings*, ed. Hans Reiss, Cambridge: Cambridge University Press, pp. 61–92.

Lessnoff, Michael (1986) *Social Contract*, New York: Macmillan.

Plato (1992) *The Republic*, trans. G.M.A. Grube with revisions by C.D.C. Reeve, Indianapolis: Hackett Publishing Company.

Scanlon, Thomas (1982) "Contractualism and Utilitarianism," in *Utilitarianism and Beyond*, eds. Amartya Sen and Bernard Williams, Cambridge: Cambridge University Press, pp. 103–28.

Skyrms, Brian (1996) *Evolution of the Social Contract*, New York: Cambridge University Press.

Vallentyne, Peter, ed. (1991) *Contractarianism and Rational Choice*, New York: Cambridge University Press.

Chapter 16

Rights

L.W. Sumner

Of all the moral concepts, rights seem most in tune with the temper of our time. At their best they evoke images of heroic struggles against oppression and discrimination. At their worst they furnish the material for lurid tabloid stories of litigious former spouses and lovers. Whatever the use to which they are put, they are ubiquitous, the global currency of moral/political argument in the new millennium. Liberal societies in particular seem replete with conflicts of rights: young against old, ethnic minority against majority, natives against foreigners, rich against poor, women against men, believers against nonbelievers, children against parents, gays against straights, employees against employers, consumers against producers, students against teachers, cyclists against drivers, pedestrians against cyclists, citizens against the police, and everyone against the state.

Love them or hate them, rights are unavoidable and no modern ethical theory seems complete without taking some account of them. It is therefore important to understand them: what they are, what their distinctive function is in our moral/political thinking, how we might distinguish reasonable from unreasonable claims of rights, and how rights might fit into the larger framework of an ethical theory. The aim of this essay is to help promote this understanding.

We can begin by trying to identify the distinctive kind of normative work rights are best equipped to do. Let us say that one part of our moral thinking has to do with the promotion of collective social goals which we deem to be valuable for their own sake: the general welfare, equality of opportunity, the eradication of poverty, bettering the lot of the worst off, or whatever. It is this part of our thinking which is well captured by the broad family of consequentialist ethical theories. On the other hand, we also tend to think that some means societies might use in order to achieve these goals are unjustifiable because they exploit or victimize

The Blackwell Guide to Ethical Theory, Second Edition. Edited by Hugh LaFollette and Ingmar Persson.

particular individuals or groups. One way of expressing this thought is to say that these parties have rights which constrain or limit the pursuit of social goals, rights which must (at least sometimes) be respected even though a valuable goal would be better promoted by ignoring or infringing them. Rights then function morally as safeguards for the position of individuals or particular groups in the face of social endeavors; in the image made famous by Ronald Dworkin (1977: xi), they can be invoked as trumps against the pursuit of collective goals. It is this part of our moral thinking which is well captured by deontological theories, and rights therefore seem most at home in such theories.

Rights impose constraints on the pursuit of collective goals. This very general characterization serves to identify in a preliminary way the moral/political function of rights, and also begins to explain their perennial appeal. But it is not yet sufficient to show how rights are distinctive or unique. Duties and obligations impose similar constraints: if I have an obligation to pay my income tax then that is what I must do even though more good would result from my donating the money to Oxfam. So what is the particular way in which rights limit our promotion of valuable states of affairs? And what exactly is the relationship between rights and duties? We need to look more closely at the anatomy of rights.

How Rights Work

A simple example will serve to get us started. Suppose that Bernard has borrowed Alice's laptop computer with the promise to return it by Tuesday, and Tuesday has arrived. Alice now has the right to have her computer returned by Bernard. Note to begin with that there are three distinct elements to this right. First, it has a *subject*: the holder or bearer of the right (in this case, Alice). Second, it has an *object*: the person against whom the right is held (in this case, Bernard). Third, it has a *content*: what it is the right to do or to have done (in this case, to have the computer returned). Every right has these three elements, though they may not always be spelled out fully in the specification of the right. The paradigm subjects of rights are persons, though nothing so far prevents them from being attributed as well to other beings, such as children, animals, corporations, collectivities, and so on. The object of a right must be an agent capable of having duties or obligations, since Alice's right that Bernard return her computer on Tuesday correlates with Bernard's obligation to return the computer on Tuesday. Since rights can be held only against agents, the class of objects of rights may be much narrower than the class of subjects. The object of Alice's right is a specific assignable person, since it is Bernard who has borrowed her computer and who is duty-bound to return it. However, the objects of a right may also be an unassignable group; some rights, such as the right not to be assaulted or killed, may hold against everyone in general.

Finally, the content of a right is always some action on the part of either the subject or the object of the right. This fact is obscured by the shorthand way in which we refer to many rights, where it may appear that the content of the right is a thing or state of affairs. We may speak, for instance, of the right to an education or to health care or to life itself. But in all such cases the full specification of the right will reveal the actions which constitute its content: that the state provide subsidized public education or health care, or that others not act in such a way as to endanger life, or whatever. The contents of many rights are intricate and complex actions on the part of (assignable or unassignable) others, which must be fully spelled out before we know exactly what the right amounts to. In the case of Alice's right the action in question is simple and specific: having her computer returned by Tuesday. Alice therefore has the right that something be done (by the person against whom her right is held). This kind of normative advantage on Alice's part is usually described as a *claim*: Alice has a claim *against Bernard* that he return her computer, which is equivalent to Bernard's duty *to Alice* to return her computer. In general, A's claim against B that B do X is logically equivalent to B's duty toward A to do X: claims and duties are in this way correlative. Claims are always of the form that something be done: the actions which make up their content must be those of another, never those of the right-holder herself. Since the content of Alice's right against Bernard has the form of a claim, we may call it a *claim-right*. Claim-rights constitute one important class of rights, exemplified primarily by contractual rights (held against assignable parties) and by rights to security of the person (held against everyone in general).

However, not all rights are claim-rights. Another example will make this clear. Alice owns her computer, which implies (among other things) that she has the right to use it (when she wants to). This right has the same subject (Alice) as her claim-right, but a different content and a different object. Its content is once again an action, but this time an action on the part of the right-holder rather than someone else: it is a right *to do* rather than a right *to have done*. The content of the right therefore does not have the form of a claim; it is common instead to refer to it as a *liberty*. To say that Alice has the liberty to use her computer is to say that she is under no obligation not to use it, or that her use of it is permissible. Actually, it is implicitly to say more than this, since Alice's right to use her computer (when she wants to) includes her right not to use it (when she does not want to). Alice therefore has two distinct liberties: to use the computer (which means that she has no duty not to use it) and not to use it (which means that she has no duty to use it). We normally treat these as the two sides of one (complex) liberty: to use or not to use the computer, as she wishes. In general, A's liberty to do X (or not) is logically equivalent to the absence of both A's duty to do X and A's duty not to do X. Alice's ownership right over the computer therefore entails her freedom to choose whether or not to use the computer; how this is to go is up to her. Since the content of her right has the form of a liberty, we may call it a *liberty-right*. Liberty-rights constitute another important class of rights,

exemplified primarily by property rights and by rights to various freedoms (of thought, belief, conscience, expression, etc.).

So far we have located a subject and a content for Alice's liberty-right, but not an object. Against whom is this right held? In the case of claim-rights the answer to this question is straightforward: whoever bears the duty which is equivalent to the claim. Because claim-rights specify obligations, and because these obligations are assigned to particular parties (or to everyone in general), claim-rights enable us to easily locate their objects. But Alice's liberty-right to use her computer involves on the face of it no claim (or duty); the liberty in question just consists in the absence of duties on Alice's part. It is therefore not so obviously held *against* anyone. And indeed, if we restrict ourselves just to its stipulated content, that is true: it is a right which imposes no duties. However, we know that property rights are typically protected by duties imposed on others: for instance, duties not to interfere with the use or enjoyment of the property in question. By virtue of her property right Alice has more than just the bare unprotected liberty to use (or not use) her computer as she pleases; this liberty is safeguarded by what H.L.A. Hart (1982: 171–3) has usefully called a "protective perimeter" of duties imposed on others. Bernard therefore (and everyone else) has the duty not to interfere with Alice's use of her computer (by stealing it, damaging it, using it without permission, etc.). We learn therefore the lesson that liberty-rights are not as simple as they seem: they involve a complex bundle of liberties (held by the subject) and duties (imposed on others). The others who bear these duties are the (implicit) objects of the right.

Even claim-rights are not as simple as they seem. Let us return to Alice's right that Bernard return her computer by Tuesday. Alice's claim against Bernard is, as we have seen, logically equivalent to Bernard's duty toward Alice. But suppose that Bernard needs the computer for an additional day and asks to return it on Wednesday instead. Alice can, of course, refuse the request and insist on the performance of Bernard's duty. But she can also agree to it, in which case she waives her right to have the computer returned by Tuesday and releases Bernard from his original obligation. She now has a new right (to have the computer returned by Wednesday) and Bernard has a new correlative obligation. In waiving her original right Alice has exercised a *power* which enables her to alter Bernard's obligation. Indeed, in entering into the agreement about the computer in the first place, both Alice and Bernard have exercised powers which result in the creation of Alice's claim-right against Bernard and Bernard's liberty-right to use Alice's computer. Contractual rights, which constitute one important class of claim-rights, therefore involve more than just claims; they also involve powers (and liberties to exercise those powers, and duties imposed on others not to interfere with those liberties, and immunities against being deprived of the powers, and so on). Even relatively simple-seeming claim-rights are therefore typically quite complex bundles of different elements. The core of the right is still a claim, but this core is surrounded by a periphery made up of other elements (claims, liberties, powers, etc.).

This periphery may be quite different for different claim-rights. Contractual rights typically confer on their subjects considerable discretion about the exercise of the right, including the power to waive it or to annul it entirely. Other claim-rights, such as the right not to be harmed or killed, may impose more limits on the subject's liberty (or power) to waive or annul the right. The full specification of a claim-right, including all of its periphery, can therefore be a very complex matter.

The same complexity, and the same relation of core to periphery, can be found in the case of liberty-rights. Alice's liberty-right to use her computer (or not, as she pleases) is not accompanied only by a protective perimeter of duties imposed on others. It also includes her power to annul her liberty to use the computer, either temporarily (by lending the computer to Bernard) or permanently (by selling it), plus her liberty to exercise this power, plus further duties imposed on others not to interfere with her exercise of this power, plus . . . Like claim-rights, liberty-rights are typically complex bundles of different elements. The core of the right (what it is a right to) is still a liberty, but it too is surrounded by a periphery made up of other elements (claims, liberties, powers, etc.).

A full exploration of the intricate anatomy of rights can be a complicated affair (see, for instance, Wellman 1985: ch. 2; Sumner 1987: ch. 2). Fortunately, we have revealed enough of this anatomy to be able to answer some of our questions about the distinctive normative function of rights. First, the relationship between rights and duties. Although these two deontological concepts are clearly connected, the connections between them are more complex than they first appear. There is a simple relationship between claims and their correlative duties: A's claim against B that B do X is logically equivalent to B's duty toward A to do X. Exclusive attention to claim-rights might lead one to think that rights are just duties seen, as it were, from the perspective of the patient rather than the agent. But this is not the case. In the first place, not all duties are relational in the sense of being owed to assignable persons. Bernard's duty to return Alice's computer has an obvious object (Alice) but my duty to pay my income tax does not: it is not clear to whom (if anyone) this duty is owed. If there are nonrelational duties then they do not correlate with any rights. More importantly, there is more to a right, even a claim-right, than just a claim against some correlative duty-bearer. Claim-rights, like liberty-rights, are typically complex clusters of different kinds of elements (duties, liberties, powers, immunities, etc.). Every such right will include some duties, either in its core or in its periphery (or both). But no right of either kind can just be reduced to a duty, or a set of duties. Rights also contain elements which are not duties, and are not definable in terms of duties. Furthermore, they have a structure, an internal logic, which is distinctively different from that of duties.

This brings us to our other question: how is it that rights impose constraints on our pursuit of goals? The complex structure of rights reveals two answers to this question. First, by containing duties imposed on their objects, rights limit the freedom of others to pursue valuable collective goals; they must (at least sometimes) fulfill their duty even when a worthwhile goal would be better promoted by not doing so. Second, by containing liberties conferred on their subjects, rights

secure the freedom of right-holders not to pursue valuable collective goals; they may (at least sometimes) choose to exercise their right even when a worthwhile goal would be better promoted by not doing so. Rights therefore impose restrictions on others (who must not promote the collectively best outcome) and confer prerogatives on their holders (who need not do so). By these means rights define protected spaces in which individuals are able to pursue their own personal projects or have their personal interests safeguarded, free from the demands of larger collective enterprises.

The two instances of the qualifier "at least sometimes" in the preceding paragraph deserve some brief attention. They signal that neither the duties which rights impose on others nor the liberties they confer on their holders need be absolute. And this brings us to a fourth dimension of a right (besides its subject, object, and content), namely its *strength*. The strength of a right is its level of resistance to rival normative considerations, such as the promotion of worthwhile goals. A right will insulate its holder to some extent against the necessity of taking these considerations into account, but it will also typically have a threshold above which they dominate or override the right. Should it turn out, for instance, that Bernard needs Alice's computer in order to arrange relief for a large-scale disaster in Africa then his duty to return it on time (and her claim that it be returned) may be overridden even if she wants the computer back. Likewise, the same degree of urgency may override her liberty-right to use the computer when she pleases. Rights raise thresholds against considerations of social utility but these thresholds are seldom insurmountable. Some particularly important rights (against torture, perhaps, or slavery, or genocide) may be absolute, but most are not.

Why Theories Need Rights

Since the normative role rights are equipped to play seems useful, even necessary, it is not surprising that most ethical theories make some effort to accommodate them. Given the currency of rights talk in moral/political argument, any theory that either ignored or rejected rights completely would risk dismissal as being hopelessly out of step with our ordinary moral thinking. Not all theories, however, are equally comfortable with rights and not all find it equally easy to take them seriously. Rather than make a positive case for a rights-friendly theory, we will proceed by examining three challenges to rights emanating from three different theoretical orientations. If these challenges can be successfully met then we will have better reason for thinking that only an ethical theory which makes room for rights will be worthy of our allegiance.

The first challenge comes from a surprising direction. At the beginning of this essay I noted that rights seem most at home in deontological theories. We should therefore be able to assume that any deontological theory will provide a hospitable environment for rights. But this is not necessarily so. Some such theories, especially

those affiliating with the Kantian or the Thomistic natural law traditions, have a decided preference for the language of duties over that of rights (see, for instance, Finnis 1980: ch. 8). Within such theories there is a tendency to treat rights as mere shadows cast by duties, so that any separate treatment of them is redundant. Now we already have the materials at hand for a response to this disparagement of rights, since we know that rights are not reducible to duties. It is true of claims that they are just (relational) duties looked at from the point of view of the patient rather than the agent, but rights are not just claims (even claim-rights are not just claims). So a theory which treats rights as just shadows cast by duties fails to understand their nature.

However, the redundancy thesis espoused by some deontological theories deserves a little more attention than this, since it enables us to say a little more about the distinctive normative role and contribution of rights. Thus far we have said that rights are complex bundles or packages of simpler constituent elements and shown how they function to constrain the pursuit of goals. But we do not yet have an adequate picture of the internal logic or rationale which unifies these diverse elements. For this we need (what we may call) a theory of the nature of rights. Two such theories have dominated the literature on rights. The *interest theory* holds that the point of rights is to protect the interests or welfare of their holders; it is this purpose which unifies the various elements making up a particular right and which explains why those elements are included and not others. Central to the interest theory is the idea of the right-holder as the beneficiary of duties imposed on others, or as the one whose interest provides the justification for imposing such duties (MacCormick 1982: ch. 8; Raz 1986: ch. 7; Lyons 1994: ch. 1; Kramer 1998). In contrast, the *will theory* holds that rights function so as to protect the freedom or autonomy of their holders. Central to this theory is the idea of the right-holder having the freedom to choose among a set of options, and of this freedom being protected by a set of duties imposed on others (Hart 1982: ch. 7; Wellman 1985: ch. 4; Sumner 1987: ch. 2; Steiner 1994: ch. 3; 1998). The main difference between the two theories lies in the emphasis which the will theory places on the right-holder's power to alter, waive, annul, or otherwise control the duties imposed by the right. It is the ability to exercise this power which gives the right-holder control over the normative relations involved in the right. On the will theory, but not on the interest theory, every right must involve some such means of control. (The distinction between these two theories must not be confused with the distinction between the two basic categories of rights: claim-rights and liberty-rights. Both theories can make sense of both kinds of rights.)

Each of these theories attempts to explain what rights are fundamentally *for* and each purports to apply across the full range of the kinds of rights we typically take ourselves (and other subjects of rights) to have. As comprehensive accounts of the nature of rights, each has its problems; fortunately, however, we need not decide which to accept. Either will suffice to show why rights are not redundant, even in a normative theory rich in duties. That rights have a distinctive normative

function is clearest on the will theory, for we have no other concept similarly dedicated to the protection of the freedom or autonomy of agents. We must be careful not to mistake the issue here. It is not whether the concept of a right might be eliminable in principle in favor of other concepts. Since rights are reducible to packages of claims, liberties, powers, and so on, we could in principle substitute these simpler concepts for the more complex concept of a right. But the result would not merely be impossibly clumsy; it would also obscure the point or rationale which binds these packages together. The idea of rights as protected choices illuminates that rationale and reveals why the concept of a right has a role to play for which we have no reasonable substitute. The will theory also makes short work of the redundancy thesis, since it requires that every right include some discretionary powers on the part of the right-holder, which means in turn that rights are not reducible to duties (not even the kinds of relational duties which are equivalent to claims).

Rights might seem in greater danger of redundancy on the interest theory, since they can function to protect interests while allowing no room for discretion or choice on the part of the right-holder. On this theory, therefore, unlike the will theory, a right could consist in just a claim, which is in turn logically equivalent to a (relational) duty. However, even here it would be a mistake to think that rights could simply be deleted from a theory of duties without loss. For the interest theory also provides an account of the point or rationale of certain kinds of duties – namely, that the justification for their imposition is to be found in a feature not of the agent or duty-bearer but of the patients whom the duty protects. Duties can have different grounds, which may focus primarily either on agent or on patient. The interest theory singles out duties whose rationale consists in protecting the interests of patients. Again it assigns to rights a point which would be lost in a theory which spoke only the language of duties.

Rights therefore can be safeguarded against redundancy in deontological theories. The second challenge, from the opposite end of the theoretical spectrum, focuses on the function of rights as constraints on the pursuit of collective goals. Consequentialist theories see the whole point of morality as consisting in the pursuit of a very abstract goal: bringing about the best overall state of affairs or making the world go as well as possible. Rights, as we have seen, are impediments to achieving this goal, since they both permit their subjects to choose nonoptimizing actions and require their objects to do so. We should not be surprised then to find that rights have been regarded with suspicion or even outright hostility by some consequentialists (Frey 1984). The consequentialist camp is internally divided on the issue of whether to make room for moral rights, some consequentialists being friendlier to this project than others (John Stuart Mill, for instance, was much more rights friendly than Jeremy Bentham). Toward the end of this essay I will explore a consequentialist strategy for not only accommodating rights but providing a foundation for them. Meanwhile, it will suffice to say that we have identified a normative role for rights – as protections of individual interests or choices against the demands of collective goals – which seems very appealing and

whose absolute exclusion from an ethical theory threatens to condemn that theory to irrelevancy. Most impartial observers, if asked to choose between the unfettered promotion of the impersonally best consequences, on the one hand, and its constraint in order to safeguard individual welfare or freedom, on the other, are likely to opt for the latter. Consequentialists may still choose to take the high road (though I will suggest later that they are mistaken to do so), but few are likely to follow them.

The third challenge to the inclusion of rights in an ethical theory is much more interesting than either of the other two. It emanates from relatively recent developments in feminist ethical theory (see, for example, Hardwig 1990; Sherwin 1992). Feminists have tended to be critical of approaches to ethical issues which are formulated primarily or exclusively in terms of rights. They see the discourse of rights as locking us into a legalistic form of moral thinking in which justice becomes the preeminent virtue. Justice may be appropriate in the public sphere, where individuals confront one another as strangers or as fellow citizens, but it is out of place in other contexts, especially in close personal relationships which thrive on values such as trust and loyalty. Rights, in their view, are adapted to a social ontology of isolated individuals indifferent or hostile to one another who need the protection of fenced-off private domains – the world of business or politics, perhaps, but not that of family or friendship. Feminists are also suspicious of the kind of autonomy whose protection is the centerpiece of the will theory of rights and which seems to promote a very masculine ideal of the rugged, self-reliant, self-defining individual with no roots and no intimate ties to others.

There are a number of important themes in this critique which we will do well to distinguish. Consider first the alleged individualism of rights. It is true that persons are usually assumed to be the paradigm holders of rights (we have operated with that assumption so far in this essay). However, there is nothing in the logic of rights which restricts them to individuals. Some rights, such as the right not to be assaulted or killed, belong to individuals simply as such, but others, such as the right to religious holidays or to services in one's native language, can be held by individuals only as members of an ethnocultural group. Even individual rights, therefore, need not define their holders as abstractions isolated from their social contexts. Furthermore, nothing seems to prevent us from taking a further step and attributing some rights (of self-determination or cultural survival, for example) to ethnocultural groups as wholes, where the right cannot be decomposed into the separate rights of the several members of the group. What is necessary in order to qualify as a potential right-holder is the possession of some value (interests on the interest theory, the capacity for choice on the will theory) which the right can function to protect. It is at the very least arguable that groups united by a common culture, language, history, or religion could satisfy the requirement of having either a collective interest or the capacity for collective choice. If so, then such groups could be the subjects of rights which will safeguard their liberties or restrict the ways in which they may be legitimately treated. Furthermore, if groups are capable of collective agency they will also be capable of serving as the objects of rights (the

bearers of the duties entailed by the right). Many rights (both claim- and liberty-rights) are held against the world at large, which still standardly means each individual member of that world considered separately. But rights can, in principle at least, also be held against groups considered collectively, and some rights (of exit from the group, for instance) appear to have this form. Therefore, if the point of the individualism critique is that rights can belong only to individuals or be held only against individuals, it is misconceived.

Neither does it appear to be true that rights presuppose a certain picture of the nature of society or of social relationships. Rights are versatile normative instruments which can be put to many different political uses. It is true that they can serve libertarians well, who dream of a suburban society of self-reliant burghers, surrounded by picket fences of liberty-rights, who have only negative duties of noninterference toward one another. But they are just as adaptable to the purposes of communitarians, socialists, egalitarians, or – dare one say it – feminists. Rights can be invoked to support a cutthroat competitive marketplace, but they can also promote an ideal of social solidarity by making it a requirement of justice that resources be allocated for the support of the needy and disadvantaged, or that discrimination on the basis of race, gender, or sexual orientation be eliminated, or that the vulnerable be protected against exploitation and oppression. Since women have historically been more likely than men to suffer from these social evils, appeals to rights and justice have been the main rhetorical weapons which they have used to better their lot (it is difficult to imagine the pro-choice movement, for instance, without such weapons). Availing themselves of the language of rights has not transformed women into isolated, rugged, masculine individuals; on the contrary, the reproductive rights which they have claimed have been deeply rooted in their identities as sole childbearers and primary childrearers.

The feminist critique, however, has another aspect to it which is more faithful to the social ideology of rights. Rights, as we have seen, impose duties and duties are normative constraints on the freedom of others – constraints whose justification lies in the protection they afford the rights-holder. The language of rights does therefore presuppose a social landscape in which interests often conflict and in which these conflicts must be managed in a principled way. Members of a world free of conflict might have no need for the protections afforded by rights. Since the public sphere is manifestly not free of conflict, rights may be conceded an appropriate, though regrettable, role in it, though even here a fixation on rights may lead us to exaggerate conflict and competition and to overlook possibilities for cooperation and reconciliation. But what place could there be for rights in the more intimate setting of a family or friendship? Personal relationships viewed through a sufficiently romantic lens might appear entirely frictionless. However, this utopian ideal is not a good fit for the daily lives of most friends, lovers, spouses, parents, and children who must also learn to manage conflicts in their personal attachments. It was a great moral and political step forward when the parties to such relationships began to be conceived as distinct individuals with a standing of their own (a process still incomplete for children, who are too often considered

even now to be the property of their parents), rather than as subordinates whose interests were submerged in those of the male head of household. To consider that these parties have rights against one another (not to be verbally, physically, sexually, or emotionally abused, for instance) is to establish certain basic expectations that every relationship should be expected to meet. The participants in any healthy, functional relationship will routinely treat one another in ways which greatly exceed this basic minimum. But that is no reason to deny that they do have such rights, and that the relationship can sink to such depths that the rights of one or more of the parties to it are being violated. Friendship may mean giving your friend more than he has a right to, but it also means not giving him less.

The valuable lesson we learn from the feminist critique is that rights do not occupy the entire moral landscape. They are specialized normative devices with a particular function, one to which they are very well adapted, but they cannot take the place of other equally important values such as loyalty, trust, and care. Nor are they a substitute for other means by which we judge personal character. If rights protect personal prerogatives, then they also protect the prerogative to behave badly – that is, in ways which, while they do not actually violate any duties or rights, are nonetheless mean-minded and selfish (Waldron 1993: ch. 3). Our moral vocabulary needs the resources to describe deficiencies of character which are compatible with the most punctilious respect for rights. Anyone who believes that human interactions require nothing more than minimal regard for the rights of others would make a very unattractive friend or spouse or neighbor – or business associate for that matter. But it is no fault of rights that the indolent or small-minded might find it convenient to think that they exhaust the requirements of virtue, and it is no solution to this problem to expel rights entirely from our moral thinking. Rights have an important, indeed indispensable, job to do within any complete and comprehensive moral theory. What the feminist critique does well to remind us is that they do not and should not stand alone.

Why Rights Need a Theory

We began by noting the extent to which rights have become the common currency of moral/political discourse and we have seen how that currency can be defended against various kinds of theoretical challenge. Ironically, however, the greatest threat to the integrity of rights discourse stems from its very popularity. It is the agility of rights, their talent for turning up on both (or all) sides of every issue, that is simultaneously their most impressive and their most troubling feature. Rival interest groups that converge on little else agree that rights are indispensable weapons in the political arena. To claim a right to something is not just to say that it would be nice to have it or generous of others to provide it: rather, one is entitled to expect or demand it, others are obliged to provide it, it would be unjust of them to deny or withhold it. Once a right has been invoked on one side of an

issue it must therefore be countered by a weapon of similar potency on the other. But then if one interest group has built its case on an appeal to rights none of its competitors can afford not to respond in kind. Like any other weapons, once they have appeared in the public arena rights claims will tend to proliferate and to escalate.

In an arms race it can be better for each side to increase its stock even though the resulting escalation will be worse for all sides. Where military weapons are concerned the increased threat is that of mutual annihilation. Where rhetorical weapons are concerned what all sides must fear is a backlash of skepticism or cynicism. An argumentative device capable of justifying anything is capable of justifying nothing. When rights claims have once been deployed on all sides of all public issues they may no longer be taken seriously as means of resolving any of those issues. Indeed, the danger is that they will no longer be taken seriously at all. Just as fiscal inflation reduces the real value of money, the inflation of rights rhetoric threatens to debase the currency.

If we once pause to reflect on the bewildering array of rights invoked in both personal and political morality then we cannot avoid asking ourselves some hard questions. Do all of these rights deserve to be taken seriously? If not, which are genuine and which are spurious? And in cases in which genuine rights conflict, which ones deserve to be taken more seriously? In order to answer questions like these we need a verification procedure for rights claims, or a criterion of authenticity for rights. As we saw earlier, the full specification of any right includes four dimensions: its subject(s), its content, its object(s), and its strength. Ideally, then, we want a criterion capable of confirming or disconfirming each of these elements in any rights claim. Nonideally, we at least need some resources to sort reasonable from unreasonable claims. But where are we to find them?

This sounds like a job for philosophers. However, not everyone working within the rights paradigm is equally helpful. Some philosophers go about their business by simply assuming a certain set of basic rights and then working out their implications for social and political arrangements. Robert Nozick, for instance, opens his most famous book with the claim that "Individuals have rights, and there are things no person or group may do to them (without violating their rights)" (Nozick 1974: ix). The rights Nozick has in mind here are (nearly absolute) property rights which stand as ethical impediments to social programs employed by the welfare state in pursuit of such goals as equalizing resources or meeting the needs of the disadvantaged. Working from this premise, Nozick devotes much ingenious argumentation to the project of working out just how much more than the bare nightwatchman state might be compatible with respect for individual rights. However, the premise itself – the assumption that individuals have just these rights and no others – is given much less attention. A more recent example of this kind of "top-down" argumentation, also within the libertarian political camp, may be found in the work of Hillel Steiner (1994). Following out the moral/political implications of libertarian premises about rights can be an illuminating process, especially when libertarians themselves disagree about these

implications. It can also serve to remind those of us of more egalitarian or social-democratic persuasion what the costs would be of failing to defend a more generous set of fundamental rights. However, ultimately a top-down methodology invites the response that it is persuasive only to the already converted, serving for the rest of us merely as a theoretically interesting exercise: yes, you have shown us where we can (and cannot) get to from here, but why should we start from *here*? If we want some means of testing the authenticity of rights claims, then we want this to apply as well to the starting points of moral/political argument, which will determine the destinations we can reach.

Other philosophers have a different way of proceeding with rights which we might call intuitionist or casuistical. The most accomplished and influential practitioner of this methodology has been Judith Jarvis Thomson (1990). Unlike those who elect to work from an assumed set of basic rights, Thomson's aim is to determine which kinds of rights we have. Furthermore, she thinks that there are some general principles which lie behind the kinds of rights we tend to attribute to ourselves and others, and she wants to determine what those principles are (1990: 1). So we seem here much closer to the ideal of a test or criterion of authenticity for rights: some alleged rights will presumably fare well in terms of these principles (whatever they turn out to be), while others will fare badly. The question then becomes one of discovering or revealing these principles. This is where Thomson's intuitionist methodology comes into play. She rests claims about the kinds of rights we have on appeals to our considered moral judgments, appeals of the sort "But surely A ought to do such and such" or "Plainly it would be wrong for B to do so and so." Because she expects general agreement with these judgments, she makes no attempt to show that they are true. Furthermore, as she herself says (1990: 4), she rests argumentative weight on them by using them to draw conclusions about people's rights. She therefore takes for granted much of the content of our ordinary morality and offers no means of confirming (or disconfirming) it. Her strategy is to argue from some (hopefully uncontroversial) fixed points in our moral thinking to implications for rights which may not (without some careful argument) seem to follow from them. All moral theorizing, she tells us (1990: 33), begins with a body of data, and her data are moral judgments which she expects her readers to accept as moral truths.

What Thomson's procedure shares with the top-down methodology of Nozick and Steiner is that certain things get taken for granted and used to support arguments for other things. However, whereas Nozick and Steiner assume some very general principles about the kinds of rights we have, the judgments Thomson takes for granted are about particular cases. Furthermore, they are not about rights but about what someone ought or ought not to do, or what it would be right or wrong to do; any claims about rights appear only as the conclusions of arguments from such judgments. Her methodology is therefore more "bottom-up" or particularist. It also more closely resembles common-law judicial reasoning which tries to work from relatively fixed points in the law to conclusions about new cases. It is particularly well suited to a certain picture about morality in general, and the

territory of rights in particular, which Thomson professes to share (1990: 33): that it does not form a system governed by a small set of very general principles. In her view, therefore, it is impossible to argue to rights from any such set of principles. Rather, again like the common law, the principles must be uncovered through the process of arguing to, and about, rights.

The results which Thomson reaches by means of her intuitionist methodology are very impressive. She is particularly good at trying to work out rigorously what we mean casually when we say, for example, that a right can be overridden or forfeited. Furthermore, the kinds of rights she takes to be genuine are, for the most part, familiar features of our (liberal) moral discourse: claims against harm or invasion by others, protected liberties, and so on. However, the argumentative structure she erects is clearly only as secure as its foundations, and those foundations consist of particular moral judgments whose truth is taken for granted. What happens if some of those judgments are disputable, or disputed? Thomson says (1990: 4, 33) that a mistake on her part about any of these judgments would be just as serious as a mistake in her reasoning from them. Are any of them mistaken? This question we could not settle without working through Thomson's arguments in detail, a task which is inappropriate for this essay. However, just as a matter of autobiography, I will say that my intuitions about the cases Thomson constructs do not always coincide with hers. Wherever that is true then the further course of her argument once again becomes merely an interesting exercise for me, on a par with the arguments of Nozick and Steiner. Furthermore, as is common in analytic ethics, many of Thomson's examples are very schematic and stripped of all social and political context. Very often my response is not so much that I agree or disagree with Thomson's assumption about the case in question, but that I want to know more, often much more, before making up my mind either way. But then many of Thomson's "data" remain question marks for me, not so much mistaken as indeterminate.

The intuitionist methodology which Thomson uses to generate a criterion of authenticity for rights is very common in analytic ethics, where appeals are constantly made to "what we believe" or "what we would say" about particular cases. In a certain respect, it is unexceptionable. Since not everything can be called into question at once, we have to assume something in order to be able to argue to any conclusions. The question is: what to assume? Thomson's implicit contention seems to be that our judgments about particular cases are more secure than any general moral principles; they therefore make safer starting points for moral argument. She makes no attempt to argue for this contention, and I can think of no way of proving (or disproving) it. She may well be right, but other methodological possibilities are equally worth exploring. So far we have considered only two: arguments from general principles about rights and arguments to rights from particular moral judgments. It is time to introduce a third: that a criterion of authenticity for rights needs the resources of a general ethical theory.

By an ethical theory I mean a relatively small, coherent set of fundamental normative principles general enough to cover collectively the whole of our moral

thinking. Since some, but not all, of that thinking involves rights, the territory of rights will form a particular subdomain in the overall landscape of a theory. The idea then is that the ultimate resource to which we appeal in order to develop a criterion of authenticity for rights is the set of basic principles in a theory. But what kind of theory? Since the options here are virtually infinite, we need to simplify the problem by focusing on a few basic types of theory. Let us assume that every moral theory has a structure or hierarchy of principles, some of which are basic (serving, in effect, as axioms) while others are derivable from them (theorems). Assume further that every such principle gives justificatory priority to a particular category of moral concepts: duties, rights, virtues, the good, and so on. Then in general a theory is X-based if its basic principles give priority to concepts from category X. In this way we can classify theories as duty-based, rights-based, virtue-based, and so on. Now let us ask the question which of these types of (foundationalist) theory is most likely to generate an operational criterion for authenticating rights.

Two kinds of theory can, I think, be excluded at the outset. On the face of it, duty-based theories might seem a hospitable environment for rights, since they are deontological right down to their foundations. However, as we noted earlier, duties form a wider category than rights and need not be patient-centered in the way that is characteristic of rights. The most basic principles in a duty-based theory – the ones which tell us what our most general duties are – may therefore refer not to some feature of moral patients (such as their welfare or autonomy) but to some feature of moral agents (such as their rationality or autonomy). Any duty-based theory with this ultimate derivation of our duties will have difficulty accommodating rights in the full sense in which they have a distinctive and ineliminable normative function. Virtue-based theories, which also tend to be agent-centered, may be excluded for the same reason. If a theory derives rights from principles about the virtues (or about the virtue of justice in particular), and if it grounds these principles in turn on an account of the good of the moral agent, then it too will lack the focus on moral patients that is the peculiar contribution of rights.

The most obvious kind of theory to provide a criterion of authenticity for rights is one which is rights-based. I will pass over the problem of how comprehensive or complete such a theory could be. After all, we know that rights define only one domain of the moral territory and are out of place in others. We might wonder, for instance, how adequate a picture of personal relationships, and of the many values which such relationships can exemplify, a purely rights-based theory could ever generate. I will also pass over the question of what the basic principles of rights in a rights-based theory might look like or how they themselves might be validated. There is a problem with the very idea of a rights-based theory which runs much deeper.

In our earlier exploration of the anatomy of rights we found their constituent elements to be such things as liberties, duties, powers, and immunities. All of these elements share one important characteristic: they are the creatures of rule systems. The parallel cases of liberties and powers will suffice to make (or remind

us of) the point. A liberty defines what is normatively permissible for an agent – what she may (is allowed to) do. A power defines what is normatively possible for an agent – what she can (is able to) do (by way of altering her own or others' normative relationships). Liberties presuppose a rule system with the triad of deontic modal concepts (required/permitted/prohibited), while powers presuppose a rule system with the triad of alethic modal concepts (necessary/possible/impossible). A rule system with deontic concepts alone is capable of generating rights on the interest theory, while a system with both sets of concepts is necessary for rights on the will theory. A legal system is the best example of a rule system with both kinds of resources, which is why it is capable of conferring rights on those subject to its jurisdiction. Those rights have legal force as long as the rules which define them, and the system itself, satisfy whatever requirements are deemed to be necessary for legal validity.

But we want a theory to support not legal rights, not any sort of merely conventional rights, but moral rights – the kinds of rights we use to criticize or justify a system of conventional rights. Legal rights presuppose a system of legal rules, so presumably moral rights must presuppose a system of moral rules. In the case of legal rules we can explain the existence of the rule system in terms of some source (a legislature or a court, for instance) which has the authority to make law. But what are the conditions for the existence of a system of moral rules (or laws)? What makes these rules moral, as opposed to conventional, is precisely the fact that they issue from no authority. But then how can we make sense of their existence? By virtue of what are they rules? And what is the source of their authority over us?

A moral theory can, I think, provide intelligible answers to these important questions. But a theory that takes rights (or, for that matter, duties) as basic, labors under a special disability since it must provide existence conditions for moral rules without recourse to any deeper principles. It is vulnerable to the accusation of hypothesizing some ghostly moral realm, the analogue of a legal system, in which moral rules somehow exist with no moral legislator. We are asked to assume that these rules are capable of imposing requirements and prohibitions, and conferring abilities and disabilities, without any of the substructure which supports the existence of a legal system. Rights and duties are of course legalistic concepts, imported into ethics from the law. Whenever such borrowing occurs the question may be raised whether the concepts make sense in the absence of the framework within which they originally developed. Now we know that sense can be made of the concept of a moral right, but only if we presuppose the background of a system of moral rules. The issue is whether a theory which treats rights as basic can make any sense of such a rule system. I cannot prove the impossibility of its doing so, but the story it would need to tell about the origin and authority of the system remain deeply mysterious.

Whether or not the foregoing considerations suffice to exclude rights-based theories as the appropriate theoretical setting for rights, there is another option worth exploring. Consequentialist theories utilize the concept of the good as their

basic moral category and combine particular instances of the good into an overall goal to be pursued (or optimized). On the face of it, a goal-based ethical theory would seem to be the least likely home for moral rights, since (as we have seen) rights serve as normative constraints on the pursuit of goals. However, appearances here may be deceiving (Sumner 1987: ch. 6). A goal may be pursued in either of two ways: directly, by just aiming at it in every instance, or indirectly, by employing some more complex motivational strategy. While some goals are best pursued directly, others are not. The goal of personal happiness will serve as an example of the latter sort. If everything you do is directly and consciously motivated by the desire to maximize your own happiness, then you will almost certainly frustrate your own aim. Your goal will be much more efficiently pursued by sometimes aiming at other things (personal relationships, for instance) which require some suppression of your self-centered fixation.

Now suppose we are talking about a very abstract moral goal such as the general welfare or equality of resources. Would such a goal be best pursued directly or indirectly? If the former, then there will indeed be no room for taking rights seriously, since they must be acknowledged as obstacles to the pursuit of the goal. But there are good reasons for thinking that, like personal happiness, moral goals are best pursued indirectly. These reasons have principally to do with the cognitive and motivational limitations of moral decision-makers, whether they be individuals or social agencies. Left with no guidance save the general exhortation to promote some abstract moral goal, most decision-makers are likely to choose counterproductive means, as the result either of deficient information or of a natural human tendency to interpret situations in one's own favor. If this hypothesis is correct, then the kinds of goals advocated by most consequentialist theories will be best pursued by accepting and internalizing a set of constraints on their direct pursuit. Since rights serve as just such constraints, then respect for (a suitably contoured set of) rights might be required by a goal-based theory. If so, then there will be room within such a theory for rights to play their characteristic normative role, while the theory's basic goal will serve as the criterion for authenticating rights. A right will count as genuine on this view just in case its recognition within some conventional rule system (formal or informal) is (or would be) morally justified, where the standard of justification is promotion of the theory's basic goal.

A goal-based theory imposes an external control on the proliferation of rights: the purpose of rights is to promote some independently defined value such as welfare or autonomy, and rights are to be recognized as legitimate only to the extent that they serve this purpose. The same basic aim will therefore also serve to demarcate the subdomain of rights, by identifying those areas of private or public life where thinking in terms of rights is inappropriate or counterproductive. A goal-based theory also has no problem accounting for the rule system necessary for making sense of rights, since the only rules it requires (or acknowledges) are ordinary conventional ones (legal and nonlegal, institutional and noninstitutional, formal and informal). A moral right on this account is a right whose recognition

in some such rule system is (or would be) morally justified – no Platonic heaven is necessary of moral rules with no moral legislator.

The most familiar form of consequentialism is of course utilitarianism, which is distinctive by virtue of its welfarist theory of the good and its aggregative procedure for combining individual welfare into a sum to be maximized. But consequentialist theories come in many different shapes with different theories of the good (both monistic and pluralistic) and different procedures (both aggregative and distributive) for defining a collective goal. For our present purposes it matters not which particular form of consequentialism we have in mind, for their common property is the priority they attach to promotion of their favored goal. Trying to fit rights into this kind of collectivist framework may seem a little like trying to square the circle, but once the air of paradox is dispelled the idea has considerable attraction. It is also the working paradigm in much judicial reasoning about rights, which often takes the form of trying to locate the appropriate balance between conflicting rights. If each of the rights in conflict (for instance, freedom of political expression versus equal respect for minorities) is intended to secure some important social goal, then striking the appropriate balance between them means drawing their boundaries in whatever way will promote the optimal trade-off between these goals. Any such approach is basically consequentialist, since it treats rights as devices for the pursuit of social goals. But it is compatible with, indeed requires, taking (the appropriate set of) rights seriously.

Other kinds of theories, such as some forms of contractualism, may share with consequentialism the virtue of controlling rights externally rather than internally. My aim here has not been to provide an exhaustive inventory of the possible theoretical settings for rights, but rather to make two tentative suggestions. The first is that the strategy of situating rights within a general ethical theory is worth exploring thoroughly before we settle for the more modest methodology of particularist intuitionism. The second is that among such theories those that are goal- rather than rights-based have a better chance both of making sense of rights and of controlling the inflation of rights claims. I have not been able to give either suggestion more than a very cursory defense, but both merit further development.

References

Dworkin, R. (1977) *Taking Rights Seriously*, Cambridge, MA: Harvard University Press.

Finnis, J. (1980) *Natural Law and Natural Rights*, Oxford: Clarendon Press.

Frey, R.G., ed. (1984) "Act-Utilitarianism, Consequentialism, and Moral Rights," in *Utility and Rights*, Minneapolis: University of Minnesota Press, pp. 61–85.

Hardwig, J. (1990) "Should Women Think in Terms of Rights?" in *Feminism and Political Theory*, ed. C. Sunstein, Chicago: University of Chicago Press, pp. 53–67.

Hart, H.L.A. (1982) *Essays on Bentham: Studies in Jurisprudence and Political Theory*, Oxford: Clarendon Press.

Kramer, M.H. (1998) "Rights without Trimmings," in *A Debate Over Rights: Philosophical Enquiries*, eds. M.H. Kramer, N.E. Simmonds, and H. Steiner, Oxford: Oxford University Press, pp. 60–100.

Lyons, D. (1994) *Rights, Welfare, and Mill's Moral Theory*, New York and Oxford: Oxford University Press.

MacCormick, N. (1982) *Legal Right and Social Democracy: Essays in Legal and Political Philosophy*, Oxford: Clarendon Press.

Nozick, R. (1974) *Anarchy, State, and Utopia*, New York: Basic Books.

Raz, J. (1986) *The Morality of Freedom*, Oxford: Clarendon Press.

Sherwin, S. (1992) *No Longer Patient: Feminist Ethics and Health Care*, Philadelphia: Temple University Press.

Steiner, H. (1994) *An Essay on Rights*, Oxford: Blackwell.

Steiner, H. (1998) "Working Rights," in *A Debate Over Rights: Philosophical Enquiries*, eds. M.H. Kramer, N.E. Simmonds, and H. Steiner, Oxford: Oxford University Press, pp. 235–302.

Sumner, L.W. (1987) *The Moral Foundation of Rights*, Oxford: Clarendon Press.

Thomson, J.J. (1990) *The Realm of Rights*, Cambridge, MA and London: Harvard University Press.

Waldron, J. (1993) *Liberal Rights: Collected Papers, 1981–1991*, Cambridge and New York: Cambridge University Press.

Wellman, C. (1985) *A Theory of Rights: Persons Under Laws, Institutions, and Morals*, Totawa, NJ: Rowman & Allanheld.

Further Reading

Freeden, M. (1991) *Rights*, Minneapolis: University of Minnesota Press.

Jones, P. (1994) *Rights*, Basingstoke, UK: Macmillan.

Lomasky, L.E. (1987) *Persons, Rights, and the Moral Community*, New York and Oxford: Oxford University Press.

Chapter 17

Libertarianism

Jan Narveson

The Theory in General

What Is Libertarianism?

Libertarianism is the view that we all have one single, general, fundamental right – the right to liberty. Rights imply duties, of course: for a certain agent, A, to have a right is for A to have a status such that other people are required to behave in certain ways towards A in the respects implied by the specific content of that right. For A to have the right to do x is for A to be such that some other person or persons is or are obligated to act in certain ways in relation to A's x-ing. So one can, substantively speaking, equivalently express the libertarian view as a *general prohibition on aggression*. An important further matter: What about the enforcement of this requirement? The libertarian principle prohibits *aggression* – not, flatly, all use or threat of force; it merely restricts it to defensive purposes.

Beyond that, we may distinguish two senses of "rights." It is plausible to say, with Mill, that *all* moral duties are enforceable, "if not by law, by the opinion of his fellow-creatures; if not by opinion, by the reproaches of his own conscience"[1] – by the tendency to disdain or enthuse, and so on. But let us distinguish rights of the kind that may be enforced *by using force* against others, and those that may not, but where we are confined to remonstrations, lookings askance, and the like. The libertarian principle concerns the former. In saying that we have a general and fundamental right to liberty, it holds that the use of force against innocent persons is wrong. Whether it also addresses, or can be made to address, the latter as well is an interesting question. Later in this essay I will suggest that it can.

The Blackwell Guide to Ethical Theory, Second Edition. Edited by Hugh LaFollette and Ingmar Persson.

But is the libertarian saying that aggression against innocents is wrong in *all conceivable circumstances*? Perhaps no theory should be held to such a standard of unqualified statement. Libertarianism is often regarded as a flat-out, uncompromising theory. Whether it permits adjustments for catastrophic situations may depend on foundational questions concerning the fundamental rationale of the theory. Certainly it is not *pacifist*. The question is whether absolutely all justified use of force is nonaggressive in the precise sense of the theory. In catastrophic circumstances, defending ourselves may be impossible without doing violence to innocents. If the libertarian right is based on considerations of self-preservation or the like, then it surely allows us to prefer preventing the heavens falling to what some consider "strict justice." Why not say, simply, that justice does not apply when the heavens are falling?

A Moral Theory – Not an Ethical Theory

Meanwhile, since nonaggression is its *only* tenet, libertarianism is apparently a very narrow view. On the face of it, it says nothing about a large range of topics that have historically fallen within the range of ethical inquiry. Most generally, there are the questions, "How shall I live?" and "What is the good life?" Libertarianism, we might say, is inherently committed to *not* answering these. Within the limits set by its stricture against aggression, it holds that we are free to choose among possible lives. The lives of the aesthete, the pleasure-seeker, the hard-driving businessperson, the teacher, the mountain-climber, the saint, are all acceptable, so far as it goes, to the libertarian. We may have our views about which lives are better, but they remain, so far as others are concerned, in the realm of advice and suggestion, not of prescriptive requirement. Whether the libertarian has anything to offer even in the way of "advice and suggestion" is a good question, which will be answered, in the affirmative, in this essay.

Liberalism and Libertarianism

Libertarianism comes on as ultra-tolerant. Indeed, it is right to regard libertarianism as an extreme form of *liberalism*, which is, I suggest, the view that the source of all relevant moral values is the individuals who are subject to its rules. Each, ultimately, is to be ruled by him- or herself. If such a view is interpreted as a philosophy of *life*, it will seem totally crazy, or totally vapid – telling us to choose, but without giving us any means of choosing. But the reply to this is simply that that is not what the theory is about. Alternatively, it could be said that the means of choosing is the individual's own inner resources: you do what *you* most prefer, all things considered. But what *do* you prefer? That is, quite strictly, *your* question; it is impossible for anyone else to make your decisions: you may turn to another

for advice, but it is you who must decide which others to turn to and whether to take the advice when offered.

Libertarianism is, then, a view about one major aspect of morality in its social sense. Of course, that aspect has always been a prominent feature of moral inquiry as well. It is reasonable to identify morality, as distinct from ethics in general, as *the set of interpersonally authoritative rules* for people in society, in groups. The widest possible group of moral agents is the group of all the agents there are, and many moral philosophers, such as Kant, have proposed theories of that scope. Libertarianism is such a theory, holding that the rule against aggression holds for all people, no matter what society they live in. That is a very ambitious claim; the libertarian bites off quite a lot, theoretically speaking.

Libertarianism as Familiar

Philosophical moral theories, in this narrower sense in which such theories are proposed sets of general rules for society, have invariably been elaborations, purifications, or more precise explications of more or less inchoate customs of the societies known to their authors. Libertarianism certainly falls within that tradition. Informal rules against killing, inflicting serious physical injuries, theft of recognized items of property, and lying are familiar from virtually all cultures, albeit with much variation in detail. In proposing as the one very general rule of interpersonal behavior that all are to refrain from aggression, whatever else, it invokes a rule that will look familiar to everyone, at least as applied to fellow members of the tribe. But, typically as philosophical moral theories go, libertarianism is in intention universal. The fact that someone is from the tribe over the hill, or the nation across the sea, is held to make no fundamental difference: we owe it to all, and they owe it to us, that our interests not be pursued by aggressive means. In the main, that prohibition itself is no great surprise; only in its refusal to accept anything except the countering of aggression as justifying the use of interpersonal force is it unique.

Persons and Self-Ownership

Aggression is acting *against persons*, and thus the vague idea that individuals are to be held "inviolate" or "sacrosanct" is readily identified with the libertarian idea – though we must be careful not to impute a religious connotation if we so characterize it. But a more general and precise analysis is certainly required. One important thrust in that direction consists in identifying the fundamental libertarian status as that of *self-ownership* – a characterization which, in borrowing a term from commerce, may raise eyebrows. But the reason for it is readily discerned. To say that someone owns something is to say that she has authority over it, that she decides about its use or disposition, that she may do as she chooses with it. If x is mine, then you can only use it with my permission – you cannot just do as you

like with it; I, on the other hand, may indeed do just that. To say, then, that a person "owns herself" is to say that it is she who decides what is to be done with that self – it is "hers" to do with or to it, or to allow or forbid others from doing with or to it, whatever can be done with or to it by that self's "owner" or others.

In saying that, we raise what is surely the main interpretive problem for this view. We may do as we like with what is ours – but, this being a universal, social doctrine, only up to the point where what we do collides with the identical rights of others. Now those rights in their turn will all devolve from the other person's authoritative relation to *his* or *her* person. But obviously it is possible for A to do what would preclude some actions that B, in turn, might want to do. In general, anyone's desire to do anything is potentially in conflict with someone else's desired course of action. Your combing your hair is incompatible with my shaving your hair all off; my going to the opera is incompatible with your burning down the opera house; and so on. Libertarianism wants to allow everyone the greatest possible freedom of action compatible with the same fundamental freedom for all others, and the question is whether that idea, as such, can generate any clear and coherent rules at all.

The idea of self-ownership helps supply the answer. Selves – persons – consist of bodies and psychologies. (Whether the latter can somehow be analyzed in terms of the former is a metaphysical question to which libertarianism need have no special answers.) Central to the human being, for purposes of ethical theory, is our command-and-control center, our "faculty of practical reason." Each person is taken to have a set of interests, desires, and presumably values (which may or may not be somehow identified with the former), which add up to a more or less coherent stock of preferences capable of feeding into practical decisions; and each person has, in addition, a stock of powers – bodily, emotional, and intellectual – which are what we immediately put in action when decisions are made; and finally, each person comes equipped with a reasoning facility and a chooser or decider, which puts into action the results of deliberations in the light of his interests and his repertoire of unilaterally actualizable capabilities. The body, being a quite well-defined physical object, especially lends itself to territorial delineation: we are to refrain both from *damaging* others' bodies and also from simply *using* them, in whatever ways it is possible for one person to use another, without the latter's consent. Minds are less easy to specify in such terms, but one person may attempt to preempt another's decision-making powers, by bullying, harassing, intimidating, and so on. But ownership of one's body – putting a moral imprimatur on the de facto power of a given mind over its associated body – is a useful starting point.

The provision *without his consent* is crucial. Jones is to be the master of Jones's being. If he wants to injure himself, that is his right; if he wants to interact with Smith in some way that Smith agrees to engage in – say, sexually, or as fellow rower in a boat race, or opponent in a wrestling match – that again is his right. A's morally certified liberty is to be as nearly as possible complete, so long as A's actions are confined to this initial "domain" of A's person, and thus it extends, in principle, to suicide or euthanasia as well. Those who would maintain the

inviolability of selves above and beyond the preferences of those selves regarding themselves abandon libertarianism for some other theory.

Libertarianism and Property Rights

Many human disputes concern the use not of our own bodies as such, but of things outside anyone's bodies: pools of oil under the surface of the earth, trees, rivers, mountainsides. Some of those things are natural, existing prior to human effort, others are modified by human activity, often beyond recognition – microcomputers do not grow on trees, and scarcely resemble anything in nature. The term "libertarianism" has recently solidified in the direction of the view that objects outside of human bodies can also be owned by individual persons. The right of private property is taken by libertarians to be a straight implication of the general liberty principle, and to be very strong indeed. With Locke – who held in the seventeenth century that a person's property cannot rightly be taken without his consent, either by private persons or by governments – the modern libertarian holds that property rights are so strong as to preclude anyone else's using what belongs to any person, for whatever purposes, not only in the case where the would-be taker's purposes are evil but also where they are as good as you like. Taking, in short, must be approved by the person taken from.

A brief explanation of this view is needed, though the subject could take us far afield. This brief explanation is as follows. Libertarianism holds that we are to be allowed to do as we wish. Only people have basic rights, and libertarians hold that they have only one basic right. Having that, what about the case in which people act in ways that happen to utilize external objects? They find things that they then use, either by simply contemplating them, as with the sunset, or as a setting for exercise, as when we walk in a forest, or by bending them to such human purposes as the satisfaction of hunger or the provision of shelter or assorted kinds of pleasure. These being things we may do, provided that we do not thereby injure others, what, then, constitutes "injury" of the relevant kind? In the case of the sunset and the walk in the forest, those uses do not compete with others, generally speaking, and in such cases there is no possessing to do, other than of one's immediate position in viewing or walking – others can usually view or walk without interference. In the rare cases where this is not so, we need rules, and they will be generally of the first-come, first-serve variety. But in the frequent cases in which the objects involved are smallish, nearby, and such that if person A uses them as person A wishes, then person B cannot use them as person B wishes, then need for a rule such as the first-possession ground of ownership is dominant.

Here there are really only two distinctive views to consider. One is the case where somebody else is already using the item in question for his purposes. It is clear that a theory forbidding interference with others' activities generally will also forbid it in the special case where that other is already using a thing. Property is *rightful* possession; but libertarianism says that *all* activities are, so far as they go,

rightful just by being activities that their agents want to engage in, provided no others are thereby injured.

The other view holds, as most contemporary writers apparently do, that when we take possession of some hitherto unused thing, we injure others by depriving them of the opportunity for future use of that item. If that were so, it is very difficult to say what the implications would be, though it is quite clear that they would not be anything like so simple as those who infer a ground for some kind of general tax seem to think (Steiner 1994). But we need not worry about the potential incoherence of the idea, for it is wrong anyway. We can interfere only with what someone is doing, not with what she is not; deprive someone only of something she *has*, not of something she *has not* – however much she might like to have it. Here, of course, we mean "having" in a sense in which that extends into the future, rather than having instantaneously. Interference, disruption, despoliation, invasion, are real relations between actual people, not phantoms of philosophical imagination. Libertarianism allows you to dream, of course, but it certainly does not give you the right that other people make your dreams come true. We must do that ourselves, and in doing so must respect others, not expect them to stand back as you help yourself to the benefits of what they have already done, or compel them to provide you with benefits you have done nothing for.

In short, the idea that there is a restriction of the kind widely quoted from Locke, that we may take from the "state of nature" so long as there is "enough and as good left for others," is a straight misreading of the liberty principle, which in Locke's version as well as mine is a pure negative principle: "No one ought to harm another in his Life, Health, Liberty or Possessions" (1980: para. 6). We do not do this when we take what no one has as yet used or even laid eyes on.

It is also true that virtually all property is *made*, in a more robust sense than merely being discovered. Humans create, remaking the world to a degree unrecognizable to Cro-Magnon man. And we spend our lives, in considerable part, in activities of exchanging. What we exchange when we exchange, we should realize, are, *always*, *services*. Those services often consist in realigning rights over things: I give you my right over this item in exchange for your letting me have the right to that ten-dollar bill. This Peruvian slum-dweller agrees to dig in this mound for a dollar a day; that movie star acts in this film, produced by this company, in exchange for a million dollars; and so on. Every person in these scenarios undertakes to improve his or her situation relative to his or her status quo ex ante, and typically succeeds in that undertaking. Insofar as the exchanges are agreed to by those party to them, who are in turn not acting fraudulently or under compulsion by any other persons, they are accepted by the liberty principle as legitimate activities.

Almost all theorists these days seem to think that there is something inherently wrong with a society of which the above descriptions would hold. Do not the high rollers *owe* the poor something more than mere recognition of the latter's liberty? Or is not the "liberty" of the slum-dweller so far from what we had in

mind in setting out to defend liberty as to be a caricature? The libertarian resolutely answers both in the negative. The descriptions above show people relating to each other, trying to make the best lives they can manage; as such we have no business picking on them, threatening them with jail if they do not hand over sums to pay for universities, slum-clearance projects or whatever; nor should we be herding the poor into the offices of bureaucrats or imposing restrictions preventing them from engaging in peaceful activities that would better their situations if they were allowed to engage in them. But to go farther on these matters would take us beyond the confines of a short essay. Suffice to say that it is not difficult to see the connection between the liberty principle and the familiar relations of "market society." Attempts to use it to support the panoply of highly interventionist programs we encounter from contemporary governments seem bound to fail.

Owning property is having the right to do whatever you want to that can be done with it, within the limits presented by other people's property rights (in themselves and in other things). This is a unitary matter; property need not be dissected into different modes or incidents, as many seem to think. In particular, there is no relevant, morally basic distinction between capitalist-type rights and primary use-type rights: initial acquirers may use their property by ploughing it up, renting it, selling or bartering it, subdividing it, establishing stocks in it, and so on, limited only by their imaginations and the interests of others. For various purposes, indeed, we can and do divide them up. Thus the landlord sells a right to use one of his flats for some months or years, and so on. The point is only that we do not need to suppose that what we basically have is only some of, or perhaps some bundle of, the "incidents" of property discerned by most writers on the subject (and attributed mostly to Hohfeld).[2]

Refinements and Queries

Fairness and Equity

Almost all theorists insist that there is a general requirement of fairness, especially in economic matters, that upsets the preceding conclusions about liberty. But libertarians can reply that considerations of fairness are relevant only when those concerned have claims, and these claims must stem from the roles they play in voluntary activities. Of course it is fair that if B has contributed x to productive activities, then his share of the gains that were the point of the activities in the first place should be proportionate to x's marginal contribution to those gains. Even there, the "should" is weak; for wages are arranged by agreement. What people have freely settled on is, basically, what they are entitled to, whether that conforms to someone's idea of their "just due" or not. But society as a whole is not a productive enterprise in the relevant sense. If it can be said to be for the "production" of anything in particular, it is peace. And there fairness tells us that

peace is to be repaid with peace, not war; to impose forcibly on some, in order to give others what they are not entitled to, hardly rates as "fair." Similarly, since people are not equal, and rarely contribute equally to production, to give each, nevertheless, an equal share in society's product would be inequitable, not equitable. For liberty to be equal – fairly distributed – is for each to supply it equally to each other in return for each getting it for him- or herself: the equality is in the amount of forcible interference by each with anyone else: *none*.

Relation to Political Philosophy

There is a general impression that libertarianism is exclusively a political theory. The impression is understandable, but wrong – clearly, the prohibition against using other people and their property is intended to apply to everyone, whether acting privately or publicly and whether what prevents them is privately wielded force or publicly wielded force. What makes it understandable is that the libertarian view as applied to the private relations of persons is so widely accepted, at least in substance, that it is taken to be obvious and, like the air around us, becomes part of the environment. Ordinary people understand that they are not to steal from anyone – though they sometimes do, anyway. The poor person does not think that he may just go ahead and help himself to the wealthy person's car or golf clubs or house; nor does the wealthy person think himself entitled to invade the shacks of the poor.

Yet there is a remarkable divergence between libertarians and others in regard to the application of these moral truisms to the actions of governments. The libertarian, uniquely, holds that it is just as wrong for governments to take people's money or expropriate their lands as for privately acting individuals or gangs to do so. Indeed, libertarians tend to regards governments as equivalent to gangs of thieves. We may think that some such gangs are more morally praiseworthy than others: some aim at and some even achieve good results, after all, and the libertarian may even agree with others about which results are to be considered "good" and which not, so far as they go. Nevertheless, the libertarian insists, to achieve these results *by those means* is wrong, just as it would be in the case of any individual. If I am collecting money to support the Harvard graduate school, I may not do so by staging a holdup, no matter how admirable we may think that institution's activities. Why, then, may the government do that? It purports to be acting on the part of the public, to be sure; but that is always false – you will never find all of those taxed supporting any government venture to the precise extent that the amount exacted from them is what they would wish to spend on that particular cause if they had their choice. And the fact that 90% of your fellow men want to spend your money that way does not make it right, any more than, to borrow from Mill again, the fact that 99% of your fellows disagree with you on some point of philosophy or aesthetics justifies them in suppressing your opinion (Mill 1910: 79).

As libertarians see it, their position is a simple matter of consistency with principles we all accept at the person-to-person level. Extending this to the most general level has one major implication: the right moral model for groups large or small is the *association*, the *voluntary* group. Associations are formed of people who share its purposes, and are willing to work with each other in the ways more or less specified or understood by the structure of the association at the time of joining. The hallmark of the free association is that if they do not like it, they can leave. Associations are typified by clubs, businesses (including their customers, who buy voluntarily), study groups, churches, and indefinitely many others. Within the association, there will often be a governing structure, to which members are likely to pay attention because they organize the activities that are the point of the association. That structure may or may not be democratic, but the individual's option to leave preserves her liberty: if she finds the association's governing body going wrong, she votes with her feet, whether or not she has a ballot.

Custom and Community

Communities and societies are not, as such, voluntary associations; their members were simply born there, or moved there with parents, or are there for other accidental reasons. The individual may be able to leave a community, but the costs of doing so are typically high. This raises a serious question, from the libertarian point of view, as to what to say about the rules of such groups. Does custom have, as Aquinas claimed, the force of law?

Here is an example. In the movie, *Zorba the Greek*, a woman has an affair with the hero, and is stoned to death as an adulterer by community members. From the libertarian point of view, prima facie, this woman has been intolerably wronged: for engaging in a purely consensual activity with another consenting individual, she suffers the penalty of death. Has she been relevantly harmed? We are given to understand that the woman in question did not question the mores; she accepted her penalty stoically, as did the hero. What are we to think? A libertarian will surely disapprove of such customs, and will think that such communities need improvement and instruction. And so, most likely, would most people of broadly liberal persuasion today – which is to say, nearly everybody. Nevertheless, it is fascinating that in typical communities around the world, their members do not see their rules as highly oppressive, or perhaps as oppressive at all. And the libertarian can and should say that those people have the right to accept such rules. The people in those communities will also, most likely, join with their fellows in teaching those customs to their children, perpetuating what we outsiders will think of as oppression. Who is right here? Or is there a right and wrong at all on such matters?

To this we can respond that aggressive interference in such communities by outsiders is not justified, but perhaps voluntary intervention, in the way of discussion and education by persons who take on the responsibility of sympathetically

involving themselves, learning the group's language and customs, and not setting themselves up as superiors, can do some good.

But this is a good point at which to make some important distinctions.

Negative and Positive Rights

First, we need to emphasize a distinction that is absolutely basic to the libertarian point of view. This is the distinction between what have come to be called "negative" and "positive" rights. That terminology has been applied, notably by Sir Isaiah Berlin, in ways that confuse the issue as much as they illuminate it, but there is in fact a straightforward, relatively simple, and familiar distinction here. A *negative* right is one which entails duties to *refrain* from certain actions, namely actions that would interfere with, impede, or render impossible the action by the right-holder to which he or she is being said to have a right. A *positive* right, by contrast, entails not only those duties, but also duties to *assist* the right-holder in doing those things, if it should happen that that person is unable to do those things on his or her own, or with the purely voluntary assistance of others. In short, the distinction is between nonhindrance and help. The distinction is readily illustrated in the case of the idea of a right to life. A murderer violates the right to life of his victim, period: the right to life is at least the negative right that others not forcibly deprive us of our lives, and the murderer does precisely that. But consider the victim in the ditch, as in the New Testament parable of the Good Samaritan. Others walk by, eyes averted; but the Good Samaritan intervenes by positive action to prevent death, tending to the victim's wounds, and/or – in an updated version – driving him to the hospital or arranging an ambulance, and so on. Now, if the victim had a positive right to life against all and sundry, then all who walk by would violate that right, even though they did not violate the negative version; its violator was only the evil person who set upon the victim in the first place. The Good Samaritan provides the help that a positive right would *require* him to do. In the libertarian's view, he goes further than he morally must.

The libertarian view, then, is that our fundamental right of liberty is negative, not positive. The reasoning behind that is straightforward. A positive right, by definition, cuts further into our liberty than the corresponding negative one: if you are *forced* to help others in need, then you do not have your choice whether to help them. Yet your not helping them does not cut into the liberty of the victims: it does not disenable them from doing whatever they can do anyway – which, to be sure, is not much. But it does not worsen their situations as compared with what they are at the time when action could take place. Instead, it merely leaves them no better off. Perhaps it will be objected that we are even *more* at liberty in the Hobbesian amoral state of nature, where anything goes and each may help himself, as Hobbes noted, even to the bodies of others. But that "may" is not a *moral* "may": in the rule-free situation, we have *no* morally protected liberty. When "anyone may do whatever he likes," no one *can* do whatever he

likes, for his neighbor may go ahead and kill him first. Morals utilize the force of the community, as it were, on behalf of its members, and the first order of morality, according to the libertarian, is mutual defense, which permits us to live our lives as we like, insofar as we can, and to seek such help as we can get from others insofar as we cannot get by on our own.

Some have regarded the positive/negative distinction as defective, even "bogus," (Shue 1985) and in any case as fundamentally insignificant. Police, for example, cost money to maintain, and yet libertarians call upon police to uphold people's negative rights, do they not? But in fact, that is a misunderstanding. Whether we *also* have a positive right to police assistance is a distinct issue from the issue of whether we have a negative right to life and property. We can have the latter without government-maintained police, or even, logically, with no police at all. At any rate, the libertarian can certainly hold that we should do without government-monopolized police, as well as without government postal services and the rest of it. Conceptually, the distinction is clear.

The second claim was made famous by James Rachels (1975), who describes cases in which the difference between killing and letting-die is all but indiscernible. But whether it is nevertheless morally insignificant is, again, a separate issue, even in his paradigm case: the uncle who fails to lift the child's head above the water, thereby not preventing its likely death, nevertheless does not *murder* the child; the one who shoves it under in order to make sure that it dies, does murder it. And, of course, in virtually all cases, the two are sharply different. You, for example, are at this very moment failing to save the lives of millions of people, any of whose lives you conceivably might be able to save – yet you are not *killing* anyone at all, never have and (I trust) never will. The thought that perhaps you deserve a jail sentence for all those omissions would be regarded by almost anyone as too absurd to bear mention.

Now: *do* we have a positive right, even to life? To take this really seriously would be to assert that those who walk by are guilty of murder, as much so as the original criminal. Very, very few people can take such a view seriously. What most of us surely think, and act on in daily life, is that helping others in severe need is a good thing to do, something we surely *ought* to be willing to do, and ought actually to do, at least when we can do it without great trouble or danger. We think that those who do go to great trouble and risk to do such things have gone beyond the call of duty, or at least that if we are to say that they have "only" done that – as they might, modestly, themselves – then it is in a sense of "duty" quite different from that in which we all have the duty to refrain from murder and theft and the rest of it in the first place. As observed above, it seems that most people, in short, are essentially libertarians in their day-to-day dealing with others.

Duties and Virtues: Charity, In Particular

The other distinction we need to make here is that between moral *requirements* of the strong type that the libertarian wants to hold us all to in regard to

nonaggression, and moral *virtues*, in regard to what goes beyond those requirements, but in a good direction. These might be called either "duties of charity," understanding that phrase to imply that charity may not be forcibly exacted from us, or "works of supererogation." In any case, we may apply the familiar notion of virtue here. Of course justice is a virtue, and a very important one – in a clear sense the fundamental and cardinal one. But there are other dispositions besides justice which can and should come in for specifically moral attention. Possession of quick reflexes is a virtue in a basketball player, but has no particular connection with morals. Charity, by contrast, is a specifically moral virtue. What makes it so? Later in this essay, I will discuss the subject of the foundations of morality, in particular as it applies to the libertarian view; but we can recognize right away why a community would do well to commend those who volunteer to assist people in need, and in general to do good works for others. Now, many good things, such as excellence in basketball playing, are not, as such, community pursuits – though hockey comes very close to being so in Canada. But everyone has a body that can be in better or worse condition, and which it is in the interest of its possessor, prima facie, that it be better rather than worse. The disposition to help it along with assorted ministrations is plainly one we all stand to benefit from if everyone has it (and if the ministrations are competently performed). It deserves, therefore, the encouragement and support of anyone. That is what singles it out as a "moral" virtue, and there is no reason at all why the libertarian cannot join with others in such recognition.

In general, then, the libertarian can perfectly well recognize that works of charity are to be commended, praised, and rewarded by people generally. Still, says the libertarian, such positive acts of doing good to others are not basically required of us in the way that forbearances from inflicting evils are required. We may properly be compelled to refrain from doing evil to others. But for not doing good, or not enough good, we may be at most criticized, perhaps shunned – but no more. In particular, says the libertarian, we may not be *taxed*; that is a compulsory extraction of what is ours.

Are there specifically libertarian virtues – virtues that libertarians especially would and should support as such? Yes: the relief of people from oppression, for example, would seem to be a specialty of the house for anyone interested in human liberty. And being interested in human liberty, generally speaking, is a specialty of the libertarian house – though by no means a monopoly of theirs. We must, of course, be very careful to distinguish *group* "liberation" from the liberalizing that the libertarian is anxious to promote. That a large group of people, living within the same boundaries on some map, should be under the thumb of one set of leaders rather than some other set is not obviously something for the libertarian to get excited about. Whether a given movement of national independence enables its citizens to be freer than before is an open question, as it stands; it will depend on conditions specific to the case, and opinions will reasonably differ.

Where libertarian opinions cannot differ, however, is on the matter of what our basic duties are. The libertarian denies that there is an enforceable duty to liberate

others from oppression. That is something we perhaps should get into, if our talents lie that way – as they usually do not. But it is clearly not something that the libertarian can coherently hold to be an enforceable duty. Here the libertarian parts company with his or her counterpart in the supposedly liberal community: proponents of government measures to promote literacy, health, income, and other aspects of welfare, often talk as if such measures were required by respect for liberty. Not so. Respect for liberty requires that we *not aggress* against others, not intervene to deprive them of what is theirs, to stymie their activities, whatever they may be. It does not, by contrast, require Beethoven to forego the frivolous activity of creating works of art that will be appreciated by a comparative fraction of the bourgeoisie and instead help free someone from the throes of oppression. If we want to do things like that, well and good – indeed, if done right, *morally* well and good, as we have seen above. But the libertarian principle calls upon us to refrain from "invading and despoiling" *any* nonaggressive person, be they currently healthy or sick, rich or poor, foreigners or next-door neighbors, and whether or not doing so might result in somebody's being freer from oppression by some other persons than he or she otherwise might be. The libertarian principle discerns a gulf between worsening and not-bettering, and insists that our fundamental requirement is to refrain from the former, whether or not we advance beyond the latter.

In all this, we should add, the libertarian continues to be *liberal*. That is, no one may impose his values on anyone else, and therefore no one's special preferences may be held up as "community" preferences unless they are literally universally shared. The preference for being healthy rather than sick is near enough to being in the latter category, but the preference for Beethoven is not. Now, many (such as the writer) who hold Beethoven in very high esteem, and who even think that there is something especially important, especially significant, especially profound about the likes of Beethoven's Quartet in C# Minor, nevertheless have no business extracting support from those who do not see it that way. So far as libertarianism is concerned, the criterion of x's being "of benefit" to person A is that person A *sees* x to be of benefit to him (or those he holds dear), even if others do not share his values. Of course, too, we are all free to steer our activities so as to benefit those whose criteria of benefit we do share. You can leave your millions to the symphony rather than the hospital for cancer patients, even though the values in virtue of which cancer is an evil are far more widespread than those that make Beethoven's music "great."

The Duty of Mutual Aid

Is this understating our moral responsibilities? Suppose there is a disaster in your community: rivers rise, leaving people homeless, their livelihoods imperiled. Does not duty call? Should not we get out there and do our bit to help people out? Of

course, we should. As, indeed, we do: in every emergency, people spring to the assistance of their imperiled fellows. This is, one might add, especially true, as a matter of document, in such individualistic, "capitalist" places as America and Canada, where the levels of mutual assistance in time of need are positively awesome. (Few will forget the response to the Southeast Asian tsunami of 2004, when response was so enormous as to create embarrassment for authorities unable to utilize it.)

I suggest that the libertarian need have no qualms about classifying mutual aid as a duty. But is it an *enforceable* duty? May we clap in jail those who do not join the lineup to build the sandbag dikes? Certainly not; and in saying this, I am sure that I say what almost all ordinary people will agree with, especially in practice. It is a point of pride and honor to devote effort to making one's community a good place to live, and helping out when help is desperately needed is an elementary point in such efforts. But to make them compulsory is wrong. And demeaning as well. How is honor due to him who toils for his neighbor, if he toils because the police await him if he does not?

When something is both a duty and yet a *nonenforceable* one, what is meant? We need to make a distinction, for as noted at the outset, we can agree with Mill that all morality is in some sense "enforceable." But let us now use a term, "reinforceable," designed to be more general than "enforceable." We *reinforce* what we think to be right by remonstrances, by excluding what we think to be offenders from our company, or by withholding certain kinds of good services from them, as well as by praising and otherwise rewarding those who perform exceptionally well in those respects. We *enforce* when we literally curtail the other person's capacity or power to do something. Incarceration and execution, and threats of same, count in this way. These are all things that the liberty principle forbids except in the specific case of violation of others' libertarian rights. But when we talk of virtues, or of supererogatory duties, we are beyond "rights" talk, and the reinforcement we apply to those we think deficient in such respects must stay within the bounds of the liberty principle. There is still considerable scope within those bounds, however. Mill's life was made uncomfortable by his friends and acquaintances because of his long-standing, uncustomarily public friendship with Helen Taylor. But on his own principles, those friends had the right to do those things, even though he thought, plausibly, that it was wrong of them to disapprove. It is in this sense that people may be said, on the libertarian (or anybody's?) view to "have a right to do wrong." Adding, as I have here, a whole category of morals on top of rights may, perhaps, redraw the territory customarily thought to be occupied by libertarian ideas. But I think not. Instead, one should think of this as clarifying its commitments. Some have understood libertarianism to be so austere a doctrine as to preclude all criticism of any and all behavior that does not, in itself, curtail the liberty of others. But that is to assimilate all criticism to that of the judge at her bench, where the sole topic for deliberation is whether the person before her is or is not to be fined, imprisoned, or hanged. But those are hardly the only moral decisions we need to make.

Children – A Special Case

From the point of view of the long-run survival of communities and, for that matter, of the human race, no one institution is more important than the family. Families generally produce and, with occasional exceptions, raise to maturity the new persons who must continually come on the scene if there are to be people in future. Those new people begin life about as helpless as one can readily imagine, and continue, though steadily maturing, to be dependent on parents, or others, for a decade or two. In early stages of childhood, human organisms are not equipped with the kind of capacities that lend themselves to sophisticated talk of rights and duties. Young children, perforce, will do more or less as they are told.

If it is less, though, then what? The libertarian, especially, has a problem about children, and from two opposite directions. On the one hand, there is a temptation to suppose that the libertarian principle should be applied to absolutely all humans, down to newborns. On the other, if it is pointed out, as is only reasonable, that children do not have the kind of facilities we had in mind in talking of a *right to liberty*, are they, then, left in the dark? For example, should we simply declare newborns to be the property of their parents, just as we declare the newly created painting of an artist to belong to its maker?

In this short exposition, my purpose is more nearly to raise than to try to resolve such issues. But a few things will be reasonably clear, at least. In the first place, the libertarian cannot accept that there is a *duty to procreate*, since that would clearly be a positive duty of the type he basically denies. Libertarianism, of course, allows them to pursue careers, or take same-sex partners or none, instead of raising families. But this still leaves us with the question of what those adults who do have children owe to them, and more generally what we all owe to all children. The situation is complicated by the fact that most people love their children and are disposed to treat them well if they can. They need no imposed duties to do that. Still, that dodges the issue. Are libertarians to claim that even small children have only the rights of noninterference that they proclaim so strongly on behalf of adults?

If we take that view, we are faced with an embarrassing consequence. For libertarianism denies that we owe assistance to anyone, as a fundamental duty, however admirable it may be to render it. But do we want to say this of parents in relation to their children? May they let them starve? But if it says they owe them more, what is that based on? Not, certainly, on a negotiated agreement between the parents and the children – the standard way in which adults come to be bound to each other. It might be suggested that the sexual acts from which children typically result somehow carry with them the obligation to support any resulting children. But just how does this act, logically disparate from the parental duty thus allegedly engendered, manage to engender it, then?

The libertarian does have some resources for staking out a credible view on this matter – much aided by the fact that for very few parents is there any issue at all,

since parents, as noted, normally come equipped with intense motivation to do well by their children. We can say, first, that while parents do, in a sense, "make" their children, they certainly do not do so in the same sense in which they make dinner, or microchips. Thus, claims to property rights in those children are rather more tenuous than in the usual cases. Second, the children in question, whatever they may be like while children, grow up to become adults, with the usual capabilities, including the capability to make things miserable for others. In the same way that people have a responsibility not to allow their property to create nuisance or danger to others, so too they have a responsibility not to bring up their children in ways that will make them a nuisance or danger to others. And thirdly, we could agree that if parents really do not want their infant children, they do not have to take them on. But suppose that other members of the community are all ready to do so? May we not insist, in self-defense and pursuance of relevant interests, that they at least allow those others to take them, rather than, say, stuffing them in the garbage?

It can be allowed that parents have the right to instruct and, within limits, to discipline their children. Modern experience suggests that the ancient method of discipline by the rod is a terrible mistake, and we may suggest that persons interested in childrearing should, and do, make an effort to circulate such knowledge. And we may surmise that the propensity to snatch misbehaving children from their parents and put them into group homes and other institutions is a mistake, and at some point a violation of the rights of both parents and children. All such options and suggestions, certainly, need to be carefully explored – libertarianism is not a completed book. Nor should the problem about children be thought uniquely embarrassing to the libertarian; no one else has better answers here, unless your criterion of a "good answer" is merely that you happen to like it.

It may be apt to add a few words on the vexed topic of abortion, for this is a matter on which professing libertarians are much split in their views. Some want to extend the libertarian rights to all humans, and take this to extend down to the genetically fundamental, thus including fetuses and even zygotes. Others, though agreeing that all humans have the libertarian rights, deny that preborns are relevantly "human." This is clearly an issue that requires moral theory at a more fundamental level. Why should we be libertarians at all? Which basic answer we give to that may also imply the right answer to the abortion issue. But the immediately preceding discussion tends, certainly, toward the "liberal" view on abortion. Mothers, obviously having rights to their bodies, can plausibly be regarded as having the upper hand as compared with beings who have as yet no "views" about anything, no values, indeed no real awareness of the world around them.

Micro-Liberty

Liberal communities specialize in tolerance. Within their capacious walls, we can expect people of all sorts to gather and, we hope, flourish. Among them will be

groups identifying with each other from tribal or religious or other ethnic backgrounds, and sometimes by common ideology. Characteristically, or perhaps always and by definition, groups of the latter kind will have customs, moral views, imposing strong duties on their members. Recall the example early on from *Zorba the Greek*. If such a community were to emigrate to North America, the practice of stoning adulteresses would quickly come under severe scrutiny. Would it still, if the surrounding community were not merely rather liberal, but outrightly libertarian? No doubt it would, at least in the sense that outsiders would be quick to spread the word that women in such communities could, if they wanted to, opt out.

But what should we say, for example, about people who marry and then take lovers outside the marriage? Should we insist that the natural right of liberty allows people to do just that, so that their complaining spouses in fact have nothing to complain of? Or, alternatively, that promises are sacred and so this must be immoral? Again, we must remember that the libertarian idea is not a theory of life, but only a theory about the proper principles for relations among people in general. Couples are not just "people in general": they have understandings, implicit and explicit, and those understandings matter a great deal to them. A liberty principle does not undermine these understandings; but it does put them in perspective. New worlds await the adventurous; but people are entitled to remain in the old ones if they so prefer. (For more, see Narveson 1999.)

Foundations: Why Liberty?

Now that we have a fair idea of what the libertarian advocates, we are in a better position to tackle the question of why we should advocate it. What is the appeal of this position – so popular, as we have seen, with people in their everyday relations, and so unpopular with governments and contemporary theorists? To raise this question is to call for exploration of the foundations of morals. There have been many views about that matter, and it may seem hubris to broach it in the few remaining pages of a medium-size essay.[3] But there is no alternative.

The very notion of "foundations" of morals is out of philosophical fashion today. That, I think, is largely because of what we can argue to be a misguided assimilation of moral theory to general epistemology and metaphysics. The latter are familiarly held to be without any particular foundations, and for plausible reasons. But morals has nothing to do with the question of the reality of the external world or whether mind is irreducible to matter. Morals takes place in a limited and wholly familiar field, the world of everyday experience. That there are some general features of people and their environments that make it sensible to accept some kinds of general rules for our mutual relations rather than others or none, is all that need be meant by the idea of foundations, in the moral area. And those who believe that there is literally no reason why we should think it wrong

to kill each other, or right to help those in need, may perhaps be talking in the misleading way of skeptics about the existence of the material world. As Hume observes, the moment the latter leave their studies, they are comfortably familiar with the ways of such material objects as their automobiles or the trees across the street, and manage to get on quite well in relating to them. We should not seek "foundations" of morals in any fancier sense than that in which we have foundations for the view that it would be nice to be able to eat this evening.

That said, let us acknowledge that philosophers tend to seek extraordinary bases for ordinary things, such as the rules of morals. Especially pertinent here are supposed "basic moral rules"; basic in the sense that they cannot be derived from or based on anything else at all. Such was the idea of intuitionism, which held that there are moral truths that one simply "sees," that are just "there." Now, it is surely true that many people will scratch their heads if you ask them why they believe that murder is wrong, or what they suppose they mean by the word "wrong" anyway. But to infer from this that there are no reasons for these homely beliefs is to point to a problem inherent in any form of moral intuitionism: the Emperor's New Clothes. People can genuinely come to wonder why murder is wrong, or even to doubt that it is. Cloaking murder with an "intuition of wrongness" will not answer them. And if intuitionism has no more than this to say to us, the thought crosses the mind that perhaps what it says is really nothing at all.

Another tendency in moral philosophy is to suppose, to put it in terms of the distinction made at the outset, that we can, as it were, go straight from ethics to morals. We will claim, perhaps, that the meaning of life is self-realization and that the reason we ought to respect other people's lives is that self-realization is a terribly important thing. Or that there is really nothing like the life of the philosopher, and in order to be one we must all make the world safe for philosophy . . . Again, we do not *answer* the question why murder is wrong in terms such as that.

Let us review the general features of social life relevant to our problem. People have preferences of great variety; among them, often, will be some ideals, and certainly some strongly held values. When they act, they attempt to bring about the objects of those preferences, as best they can. Were there only one individual in the world, there would be no problem of morals, though certainly she would be faced with the questions of general ethics – of trying to decide what she should devote her life to; but basically, that person's problem would be just to do the best she could in the face of a not very helpful natural environment, and that would be that. But of course, we are not in that situation. We live in a social world, encountering many others in daily life. Those people, unlike trees or rocks, have minds of their own, making decisions and choices, pursuing their various values. Some of those values will be similar to ones we have, but even the similar ones have an important potential for bringing us into conflict. And they have another important potential – to help promote projects we do value. Rationally, we want to avoid the former and promote the latter.

Why do we care about our own liberty? The obvious reason is that liberty, in social terms, is the absence of acts by others preventing us from doing what we

are interested in doing – as any of them could do. Given liberty, we are on our own steam, at least; lacking it, we cannot get where we want to go.

Then there is the potential for help at the hands of others. This potential, in principle, can be harnessed in two ways. One is by using force – enslaving, coercing them into providing us with the desired benefits. The other is by cooperative methods. These are the relations to others wherein both the other and the agent benefit as a result. A does something for B, B does something for A. Cooperative relations have many nice properties. Most important is that they are good for us both – by definition. A close second is that they reduce the other person's motivation to make life miserable for his partner. If you benefit yourself in the process of benefiting me, you have reason to continue, rather than to turn to assault and domination.

But it is pretty easy for people, even when in potentially cooperative relation to each other, to be tempted instead to take advantage of each other. When A's back is turned, perhaps B should attack, thus gaining the benefits already got from A, while eliminating the onerous need to do his share. This, it is thought, is how *egoists* would behave. Yet if both made a habit of it, there would be problems – to put it mildly. Two persons disposed to take advantage of each other at every opportunity are not two winners, but two losers. Devoting our time to fighting or, in suspicious efforts to cover ourselves, building defenses, and so on, is not going to get a great deal done.

When we consider one-on-one relations with others, it is not difficult to realize the potential value of rules calling for cooperation and forbidding tendencies to try to dominate and extort. And if we broaden the net a bit, another large factor comes into prominent view: the differing abilities and interests of people are conducive to specialization of many sorts, increasing the likelihood that any given person will be able to find others who will cater to his or her peculiar interests. Moreover, some among those many will be brilliant, ingenious, creative, industrious, enterprising, and so on. Such people will, if not prevented, come up with useful ideas about how to do things that might very well improve our lives.

The short of it – there is no room for the "long" – is that cooperation fits very well into anyone's profile as a person. Almost all of us have considerable abilities that we can, if allowed, bring into play to realize what we want in life, almost whatever it may be. All of us differ at least somewhat, and many a lot. How do we do best insofar as general rules of conduct can affect the issue?

The "rule" that *others* are to devote themselves to *our* welfare has an initial appeal, but one that withers under even fairly superficial analysis. First, and especially, why should they do so? That is to say, why *would* they do so, in view of the fact that they do not carry in their genes any special affection for us? But even that is misleading. For many of them do carry in their genes general affection for us, enough that, on a decent day, they are quite likely to be helpful where help is relatively easy to render and can do a lot of good. But what nonlibertarians propose is a rule of the heavy-weapons type: we should stand ready to enforce the rule of helping others with the threat of jail and the like. Such a rule, for this purpose,

has two major disadvantages. For one thing, its administrative costs are high: jails are expensive, not to mention the legal and other apparatus that goes with them. By contrast, voluntary cooperation carries its own fuel: it is largely self-enforcing, as each has motivation to participate so long as the other does, and both realize that cooperation does not end sharply at noon but goes on the next day and the day after. When each is motivated to cooperation by interests he or she actually has, there is little or no need for outsiders to participate. Coercive methods, on the other hand, are necessarily inefficient. Why go for relations in which some gain only at the expense of others, when instead we can have relations in which one person gains only when at least some others do too?

But further: we do not want the help of others to be a matter of sheer happenstance. It is in everyone's interest that everyone has a sense of duty toward others, to help when they easily can. The ideal compromise is to make this a matter of noncompulsive obligation, as discussed before. We will praise those who are helpful, and will downrate those who are not when they readily could be. The libertarian's objection to compulsion holds here as well; but after all, the point of morality is to enable all to live well, and in a limited but important class of cases, the help of our fellows when in urgent need may be essential to our doing so. All are to be *allowed* to live as they please; but all are to be *encouraged* to charity and good service to their fellows.

It is obvious why each of us wants to be free of real interference, so far as we can. It does not follow directly that we would do well to make it a right. But the cost of accepting a general right of liberty, in which each is allowed to do as he or she prefers, with the minimal constraint compatible with everyone's being able to exercise this right, is low. Positive rights would add to the burden, without providing compensating advantages for all: not only are we to desist from invading and despoiling but we *also* have to devote some or a good deal of our energies to helping to realize the aims of others, even when we do not share those aims – as we typically do not. Why would we do that? Not for the sake of the gains it enables us to have – for if we all could gain, we would not require anyone to stand over us with the power to jail us for nonassistance. And if only most of us could gain, why would we not form an association of like-minded persons, who would commit themselves to help fellow members without imposing extra burdens on others outside it? Why would those potentially outside it accept a general rule imposing those pointless and irksome burdens, as they surely are from their point of view?

The liberty principle accords quite well with the core of ordinary morality as practiced by most people most of the time. But that, as such, is not what recommends it. Instead, what recommends it is the same thing that recommends it to those ordinary people themselves: it is the principle that can be most expected, among possible principles, to enable each of us to live the better lives that we envisage to be achievable, given our various interests. There are reasons explaining why few societies have, nevertheless, looked very much like libertarian societies, despite the general acceptance of the liberty principle among people in their daily

lives. Those reasons are based on the logic of coalitions – of ganging up on people to exact short-term gains. In the longer run, though, we all suffer at the hands of such systems. But that is a subject for another discussion.

Notes

1 J.S. Mill (1910: 45). For example, "We do not call anything wrong, unless we mean to imply that a person ought to be punished in some way or other for doing it; if not by law, by the opinion of his fellow-creatures; if not by opinion, by the reproaches of his own conscience. This seems the real turning point of the distinction between morality and simple expediency. It is a part of the notion of Duty in every one of its forms, that a person may rightfully be compelled to fulfil it."

2 Hohfeld, Wesley (1919/1920). Hohfeld's distinctions are discussed in innumerable sources. See, for example, the Wikipedia article on Hohfeld, http://en.wikipedia.org/wiki/Wesley_Newcomb_Hohfeld (accessed November 15, 2012).

3 For a more extensive discussion along this line, see Narveson (2010).

References

Hohfeld, Wesley (1919/1920) *Fundamental Legal Conceptions as Applied in Judicial Reasoning and Other Legal Essays*, New Haven, CT: Yale University Press.

Locke, J. (1980) *Second Treatise of Civil Government*, ed. C.B. McPherson, Indianapolis and Cambridge, Hackett Publishing Company.

Mill, John Stuart (1910) *Utilitarianism, Liberty and Representative Government*, New York: E.P. Dutton.

Narveson, Jan (1999) "Abortion," in *Moral Matters*, 2nd edn, Peterborough, Ontario: Broadview Press, pp. 157–80.

Narveson, Jan (2010) *This is Ethical Theory*, Chicago: Open Court.

Rachels, R. (1975) "Active and Passive Euthanasia," *New England Journal of Medicine* 292: 78–80. Widely reprinted, in Jan Narveson (1983) *Moral Issues*, Toronto and New York: Oxford University Press.

Shue, Henry (1985) "The Bogus Distinction – 'Negative' and 'Positive' Rights," in *Making Ethical Decisions*, N. Bowie, New York: McGraw-Hill, pp. 223–31.

Steiner, Hillel (1994) *An Essay on Rights*, Oxford: Blackwell.

Chapter 18

Virtue Ethics

Michael Slote

Virtue ethics was the dominant approach to ethics in the world of (Western) classical antiquity, but for reasons to be discussed in what follows, its influence waned during most of the modern era, and it is only in recent decades that it has revived as a major approach to moral theory. I am going to talk about what virtue ethics *is*, about the different forms virtue ethics has taken or currently embodies, and about the philosophical issues that those favoring virtue ethics face in their attempts to show its superiority to other (contemporary) approaches.

The Nature and Variety of Virtue Ethics

Many philosophers have spoken about virtue and the virtues without, in the contemporary sense, counting as virtue ethicists (or "virtue theorists"). For example, Kant (1964) has a "doctrine of virtue," an account of moral virtue that flows out of what he has to say about right and wrong action, and Rawls, too, in *A Theory of Justice* (Rawls 1971) has an account of moral worth or virtue that derives from and complements what he has to say about the principles of justice. But any view that treats virtue as simply one part of a moral theory does not count as virtue ethics or virtue theory. Virtue ethics seeks an account of virtue that is self-standing and fundamental, rather than derivative or complementary to other ethical ideas, and we might then say (slightly altering a suggestion of Roger Crisp's) that Kant and Rawls have theories *of* virtue, but are not necessarily proponents of *virtue ethics* or *virtue theory*.

The Blackwell Guide to Ethical Theory, Second Edition. Edited by Hugh LaFollette and Ingmar Persson.

We can say more about what such virtue ethics is by contrasting it with other well-known schools of moral/ethical thought. (It is probably impossible to give an airtight definition here – just as it is with everything else philosophers talk about.) Both intuitionistic and Kantian deontology tend to treat the solution of moral problems or difficulties as dependent on finding rationally valid rules or principles for dealing with them. Utilitarian and other consequentialists treat the moral evaluation of human actions, motives, or institutions as depending fundamentally on actual or likely consequences or results. But virtue ethics treats neither consequences nor general principles as central to its and our understanding of morality. The emphasis, rather, is on inner character and motivation, and this is as true of recent virtue ethics as it was of Greek and Roman virtue ethics. The virtue ethicist holds, in other words, that if we want to decide what is right or ethical, we need to appeal to facts about human character, human motivation, and human virtues and vices in order to ground or justify our conclusions. And rules/principles and the general results of our actions are relevant to moral evaluation only to the extent that they illustrate or exemplify certain valued or disvalued inner/psychological dispositions.

Plato and Aristotle were both virtue ethicists in this sense (as were the ancient Stoics and Epicureans and, as we shall be seeing, various Asian thinkers as well). Plato in the *Republic* says virtue consists in a certain harmony or beauty of the soul, and actions are considered ethically better or worse to the extent they sustain or fail to sustain such inner harmony/beauty. Ordinary moral rules are treated as a secondary (theoretical) issue, and the question of external results (e.g., whether one is making people more contented with their lives) is left largely to one side. Similarly, Aristotle denies that the moral life can be governed by rules or general principles, and the criterion of right or noble action is the rational ethical perceptions of the virtuous individual in situations of ethical choice.

Perhaps the most important modern proponent of (something like) virtue ethics was David Hume, who argued that actions have moral value only insofar as they express good underlying motives. But Hume also placed more emphasis on good results or consequences than Plato and Aristotle did, so he is a precursor of contemporary consequentialism at the same time that he is (for that very reason) a less than pure practitioner of virtue ethics as it was described above. But there are a whole host of other ancient and modern virtue ethicists whom we will not have time to discuss in this brief essay, and it is perhaps most important at this point to say some things about the contrast between recent virtue ethics and historically earlier forms.

Virtue ethics began reviving in the West (it has never lost its influence in Asian ethical thought) under the influence of an article by G.E.M. Anscombe entitled "Modern Moral Philosophy" (Anscombe 1958). In that essay, Anscombe took strong issue with the Kantian and consequentialist approaches to ethical theory that had dominated English-speaking moral philosophy during the modern period. (There is no space to discuss her specific objections.) She also argued that we need a more Aristotelian approach to human moral psychology before ethics can go

forward in a theoretically acceptable fashion. And many ethicists were inspired/influenced by that article to look again at Aristotle as a potential model for how to do ethics. In its earliest days or decades, reviving virtue ethics was (therefore) almost exclusively Aristotelian in character, but eventually it came to be recognized that Hume, Plato, the Stoics, and even Nietzsche might be looked to for virtue-ethical inspiration, and nowadays (and not surprisingly, given the increasing political/economic power of China) Confucianism is also seen by many ethicists as a possible model (or cluster of models) for contemporary virtue-ethical thought. So as virtue ethics has become ensconced as one of the three major contemporary approaches to ethics (alongside Kantian ethics and consequentialism), the virtue ethics movement has lost its originally monolithic Aristotelian character, and those modeling their virtue-ethical thinking on one or another historical (or contemporary) influence have to defend/justify their particular ideas against other forms of virtue ethics and not just against consequentialism and Kantianism. (In this respect, the situation of consequentialists and of Kantians is pretty much symmetrical with that of the virtue ethicists. Consequentialists, for example, have to decide between utilitarian and nonutilitarian, and between direct or indirect, forms of their approach.)

In the light of these differences, I think it might be useful if I made some somewhat general distinctions (not tied to any particular philosophers) among various kinds of virtue ethics. And perhaps I should begin by pointing out that all the virtue ethicists of classical antiquity were eudaimonists and that this is decidedly not true of all modern or contemporary virtue ethicists (much less of Kantians and consequentialists). Eudaimonism is, roughly, the doctrine that some character trait counts validly as a virtue only if someone possessing the trait gains some sort of personal advantage by doing so. The eudaimonist holds that genuine virtue has to be by and large in the interest of the virtuous individual, and modern-day moral thinkers – whether virtue ethicists or not – often do not hold such a view. The consequentialist says that a morally good person is someone whose actions or inner character has a positive effect on overall human welfare, and sometimes such goodness requires an individual to *sacrifice* their own welfare for the sake of others (or of humanity generally). Similarly, Kant does not make his arguments regarding moral obligation and moral goodness depend on the assumption that a morally good person will inevitably or automatically be better off as a result of their goodness. And many recent or contemporary virtue ethicists also regard it as possible that the virtuous person should benefit others to or at their own cost: this is true of the late-nineteenth-century figure James Martineau (1891), but also, though in different degrees, of recent virtue ethicists like Rosalind Hursthouse (1999) and myself (2010).

One reason why modern-day ethicists, including many virtue ethicists, seem much less wedded to eudaimonism than the Greek and Roman philosophers were has to do, I think, with the influence of Christianity. Self-sacrifice was not a major ethical concept in the world of classical antiquity, but the example of Jesus's sacrificing himself by suffering and dying for our sins has influenced modern moral

thought in a profound manner, and the relatively secular philosophers of the modern era have (consequently) seemed willing to entertain the possibility or occasional necessity of virtuous self-sacrifice in a way that never or almost never seems to have occurred to ancient, pre-Christian philosophers. (I say "almost never" because there is a controversial passage in Plato's *Republic* that seems to suggest that the virtuous philosopher will sacrifice his greater happiness as a pure contemplator of Reality in order to serve the greater good of the State in the role of philosopher king. But by and large Plato accepts and argues for eudaimonism.)

Another important distinction among virtue ethicists concerns the validity of ethical theory as such. Anscombe objected to consequentialism and Kantian ethics on somewhat theoretical grounds, but many of those who were influenced by her article thought that the chief problem with the reigning views was that they *were theoretical*. Virtue-minded philosophers like Annette Baier (1985), John McDowell (1979), and Bernard Williams (1985) argued that Kantianism and consequentialism neglect vast areas of human life (e.g., have little to say about friendship) and/or oversimplify the vast riches, complexities, and difficulties of the moral life in the name of theoretical unification. Such antitheorists held that theoretical models of or in ethics in effect treat ethics on a par with scientific theorizing and fail to recognize the essentially practical and normative purposes that precisely distinguish ethics from any form of science.

However, other (somewhat later) recent virtue ethicists have noted that Aristotle himself was an ethical theorist and, drawing inspiration perhaps from Thomas Kuhn's idea (Kuhn 1962) that it takes a theory to beat a theory, have suggested that the virtue-ethical critique of Kantian and consequentialist moralities can be fully successful only if virtue ethics can offer an overall and theoretical conception of ethics that is superior to what these other approaches have to offer. (Both Rosalind Hursthouse and I have said things in this direction.) However (running somewhat against the grain of what has just been described), some fairly recent virtue ethicists have questioned whether virtue ethics needs to be practical in the sense of offering guidance with regard to particular actions and choices. Following the Victorian thinker Leslie Stephen (1882), some virtue ethicists like Edmund Pincoffs (1986) have held that virtue ethics should frame a conception of what it is to be a virtuous individual, or of what traits count as virtues, and leave questions of what to do and how to act to the discretion of the situated virtuous individual. However, most recent virtuous ethicists are convinced that virtue ethics needs to be able to say something practically useful and illuminating about right and wrong actions, and it is frequently maintained (e.g., by Hursthouse) that virtue ethics can offer useful moral rules about how to act as long as it is understood that such rules derive their moral force and validity from considerations more basically having to do with character or good motives as such and conceived independently of the rules. (Consequentialism typically has a similar, secondary or derivative place for familiar moral rules.)

Another important distinction among virtue-ethical approaches concerns ethical egoism. Is the fundamental concern of morality one's own self-interest or is a basic

concern for others built into our (virtue-ethical) understanding of the moral life? In ancient times, there was a tendency for all ethicists to be ethical egoists, and most scholars would grant that Stoicism and Epicureanism constitute egoistic forms of virtue ethics. But Aristotle (and arguably also Plato) is different. Aristotle is a eudaimonist who holds that virtue is essential to and helps to constitute our happiness, our well-being, but he does not accept the egoistic idea that we should always be motivated by considerations of self-interest, that we should always attempt to do what is good or best for ourselves. Aristotle holds that the warrior who gives his life for his country has a better though shorter life than the ignoble coward who flees from the enemy, but he seems to hold that the virtuous courageous soldier is motivated by the desire to protect his country and/or the desire to do what is right and noble, rather than by the consideration that he will have a better life if he fights and dies than he would if he ran away and lived a subsequent life of shame and dishonor. So Aristotle is a eudaimonist, but not an ethical egoist, and by and large most modern-day virtue ethicists are also not ethical egoists. To be sure, Nietzsche is often considered to be a virtue ethicist, and he recommends that the individual act from power and do what increases his power vis-à-vis others. But his views are not necessarily egoistic in the strictest sense, because he praises and even advocates a kind of "Godfather generosity" that causes or inclines the powerful individual to give to weaker others out of a sense of his or her own power and self-sufficiency. Such generosity does not aim *at* the giver's own good, but comes *out of* a sense of the giver's superabundant well-being; so even if Nietzsche has nothing nice to say about compassion, pity, guilt, and conscience, he is nonetheless not an ethical egoist.

Of course, those, like David Hume, who emphasize and praise compassion and benevolence also are not egoists, and neither does Hume seem to be any sort of eudaimonist (though he was not very clear about this issue). But the contrast between Nietzsche and Hume as virtue ethicists should also put us in mind of another important, even definitive, distinction within the field of virtue ethics. Hume was a moral sentimentalist who advocated benevolence, compassion, and the like and also thought that our praise of these qualities was not grounded in reason, but rather in the human capacity for sympathetic feeling, in sentiment. But many or most other virtue ethicists have been rationalists and have held that both moral motivation and the discernment of moral truth are grounded in our capacity for (practical) rationality. Such ethical rationalism predominated in the ancient (Western) world; but nowadays those virtue ethicists (and allies of virtue ethicists) who look to Hume and the other moral sentimentalists for ideas and inspiration are sentimentalists, whereas those who follow Plato, Aristotle, or the Stoics (there are not many Nietzscheans or Epicureans) are rationalists in the same basic sense in which Kant is a rationalist.

And the divide or distinction between rationalists and sentimentalists within the field or school of virtue ethics takes on added significance when one realizes that this distinction is relevant not only to historical and contemporary Western views, but also to the traditions of Confucianism. So let me now try to say something

more about how the distinction between sentimentalism and rationalism applies within Western virtue ethics, both historically and today, and then show the relevance of that distinction to our understanding of Confucian thought.

As I indicated above, there were no sentimentalists in pagan Greece and Rome. Everyone who had positive ideas on the subject thought that our beliefs about right and wrong, good and bad needed to be grounded in reason or Reason. Plato held, for example, that our understanding of moral truth rested on our ability to rationally apprehend the Forms that he posited as the highest or only reality. Our knowledge of what is virtuous or vicious in human life therefore depended on our capacity for reason or rationality, and Plato also thought that this knowledge, once or fully attained, automatically translated into moral motivation. To accurately apprehend the Form of the Good or of Justice is automatically to be moved to act in the light of that apprehension, to do what our knowledge of the Good or the Just tells us is good or just. So for Plato at least moral knowledge was tantamount to moral motivation – just as one might think, extra-ethically, that a vivid and accurate apprehension of beauty could automatically translate into a desire to prolong and preserve the experience one is having, one's aesthetic commerce or connection with the beautiful.

Aristotle held a teleological view of everything in the universe and saw our capacity for rational activity as the defining essence and purpose of human beings. He also argued that human well-being or eudaimonia is tied to the fulfillment or actualization of that essential purpose (over a long lifetime). Aristotle's account of the role of reason in ethical knowledge was much less otherworldly than Plato's. Our appreciation of moral truths that are relevant to our lives did not have to appeal to the extra-sensory Forms that Plato posited, but was ensured and effected, rather, by an inculcated and cultivated rational disposition, independently of pregiven moral rules, to appreciate what any given situation ethically calls for or demands. Aristotle likened this appreciation to perception and to seeing in particular, but he held that the perception and seeing is of a strictly rational kind and is at best only analogous to what happens in sense perception. The particularistic (non-rule-guided) rationalistic apprehension of what is morally required in a given situation is also treated by Aristotle on analogy with aesthetic appreciation and connoisseurship (which also do not appeal to or depend on rules), and on some interpretations (see McDowell 1979) the proper apprehension of situational ethical facts automatically translates into, or gives rise to, appropriate ethical motivation. (However, some interpreters think Aristotle needs to appeal to desires that are independent of our knowledge of ethical facts in order to explain why we are moved to do what is right or noble.)

In any event, as with Kant much later, there is a certain tendency in Plato and Aristotle to explain and justify both moral knowledge and moral motivation/action in terms of our rational or reasoning powers, but the gain in unity here is not necessarily to the advantage of such views if the grounding idea that the apprehension of moral facts or truths is automatically motivating seems implausible. J.L. Mackie (1977) has described this idea as "queer," and from a certain

perspective it does seem queer. How can sheer knowledge of (moral) truth or facts motivate us to *act*? However, the idea that the appreciation of real beauty automatically moves us in certain ways may take the edge off the queerness: if this can happen in the realm of beauty, why not in the realm of the good and the right as well. And a further analogy also makes me think that the idea of motivating knowledge of the truth (what Mackie calls "objective prescriptivity") is less odd than one might initially think. As my colleague Elijah Chudnoff has pointed out to me, our recognition of the validity of a certain form of inference is inseparable from a disposition (not always effective) to reason in accordance with that form of inference. And why should not what thus holds in epistemology also hold in other areas? The idea that knowledge can in principle motivate is not implausible as such, and in that case its application to ethical theory is far from ruled out in advance. So rationalistic virtue ethics and Kantianism are not necessarily criticizable for the close connection they argue for between moral knowledge and moral motivation (dispositions to moral action). And perhaps more interestingly, moral sentimentalism may also (as we shall see in a moment) offer philosophical reasons for tying truths (it holds to be) delivered via feeling and sentiment to appropriate moral motivation. But the real question for sentimentalism, actually, is whether it can or should want to accommodate and justify the idea of moral truth in the first place. Let me explain.

The virtue ethicist who looks to Humean-type sentimentalism for ideas and inspiration will want to say that "feelingful" motives like benevolence and compassion mark a principal part of morality – in other words being morally decent or morally good is largely a matter of acting from the just-mentioned motives or sentiments and without reliance of rules or principles that might have to be justified in terms of reason or rationality. The person who saves someone in danger of drowning does not (necessarily) appeal to moral rules or rely on a sense of obligation or duty before they act: they feel compassion and act primarily on the basis of that "feelingful" motive. So Hume (1739/1978) and other sentimentalists think of the moral life as to a considerable extent a matter of acting from feeling rather than from reason or rational considerations, but the question then arises as to how explicit moral thought can arise out of such naturally good and presupposed human motivational tendencies. The person who compassionately saves someone from drowning or from some form of suffering need not be thinking in terms of obligation, duty, or virtue, but sometimes we *do* think and *need* to think about our moral obligations, and the question then arises as to what makes such thinking possible. For the rationalist, such thinking is strictly rational, the rational apprehension, for example, of the truth of certain principles or of situational/particularistic ethical facts/requirements.

But the sentimentalist cannot and does not want to say this kind of thing. For the consistent or thoroughgoing sentimentalist our capacity for *talking and thinking about* morality depends as much on our sentimental nature and capacities as our capacity to *do* good or obligatory things at the behest of our natural benevolent motivations. (There is an issue about how or whether these motives need to be

cultivated, but it is probably best to leave that issue to one side here.) In the first instance this means, or at least suggests, that talk about what is right or wrong expresses certain sympathetic or benevolent feelings on our part. In other words, it may be the case that when we say that some action is or would be wrong, we are expressing our sympathetic benevolent preference that such an action should not be done, our concern, say, for the person or persons who we think would be the victim(s) of that action. And a similar story could be told about positive judgments of moral goodness: (roughly) we call something, some action, morally good if our benevolence wants to see it occur. On such a view moral utterances or beliefs express our (complex) sympathetic concern for other people (or animals), and one finds something like this metaethical view of moral utterances in Hume and in other (metaethical) sentimentalists like Charles Stevenson (1944) and Simon Blackburn (1984). But I believe, and others have claimed, that such a view of what is happening when we say that something is right or wrong or good or bad creates problems for sentimentalism – problems that rationalism does not face (and is intended to evade or obviate).

The problem is that when we say that something – like torturing people for the fun of it – is morally wrong, we typically or often feel – at least before we start doing philosophy – that we are saying something true or valid: that such torture is, really is, wrong. Rationalism allows us to preserve this initial intuition by telling us that we learn of the objective wrongness of torturing people for fun via the use of reason. But the sentimentalist tells us or may tell us that when we say torturing (for fun) is wrong, we are expressing our benevolent motivation toward those who might be tortured and our displeasure, also in the light of our sympathetic benevolence, with the person who wants to torture some person or persons. And if that is all that we are doing, then we are just expressing our feelings *rather than making any sort of statement that could lay claim to objective (moral) validity.* In a nutshell, Hume is often said to be an expressivist or emotivist or subjectivist about moral claims or utterances and to be committed, therefore, to denying what most of us antecedently think and have motivation to hold on to, namely, that moral utterances like the one about torture have some claim to being objectively true. So virtue-ethical sentimentalism has, on the face of it, certain seeming disadvantages in comparison with rationalist forms of virtue ethics or of ethics more generally, and in the second part of this essay, I want to consider whether sentimentalism can offer an answer to this kind of criticism. For the moment, though, it is enough to have noted that like rationalism, sentimentalism has its own distinctive account of moral motivation and of the meaning of moral claims or utterances. The opposition between these two virtue-ethical approaches is thus very deeply drawn, and in what follows I want to consider the present-day plausibility and prospects of these two very different ways of conceiving and approaching ethical questions. But before we do that, let me indicate to you how the division between sentimentalism and rationalism carries over to our understanding of the Confucian tradition (and to some extent, though we shall speak less about this, to other historical trends and tendencies within Eastern/Asian thought).

The two most influential figures in Confucianism are Confucius (Kongzi) and Mencius (Mengzi) – the only two thinkers in that tradition whose names have romanized versions. And there has in the West been a tradition of seeing these two thinkers and many of their followers as Aristotle-like virtue ethicists. Now some interpreters think of Confucianism as more like "role ethics" than like virtue ethics; but those who see Confucianism as a kind of virtue ethics can point to a number of similarities between Confucianism and Western virtue ethics, most notably, perhaps, the great emphasis on virtue and being virtuous rather than on good consequences and on fundamental rules or principles. (The ancient Chinese thinker Mozi was the first consequentialist to live on this planet, but Confucianism routed Mohism at a fairly early date, and that tradition simply died out in China.) But the comparison between the Confucians and Aristotle has a somewhat problematic character, because although Aristotle stressed the foundational importance of reason, the thinkers of the Confucian tradition do not rigidly or rigorously distinguish between reason and sentiment (their central concept of *xin* is often translated as "heart-mind"); and that already indicates that there may be problems with calling Confucius a rationalist rather than a sentimentalist and thus with comparing him and other Confucians (or neo-Confucians) *with Aristotle rather than with Hume*.

What may help us gain focus on this whole issue is a consideration of how the central Confucian virtue-ethical notion of *ren* seems to shift its implications or emphasis between the Confucian *Analects* (1979) and Mencius's later work, the *Mencius* (1970). In the *Analects*, *ren* is used in a fairly general way to refer to virtue on the whole or as a whole. But in the *Mencius* the same term is used in a narrower fashion to refer to compassion and/or benevolence. Between Confucius and Mencius a certain shift toward a distinctively sentimentalist understanding of virtue seems to have occurred, and Mencius's well-known example of how compassion is stirred by seeing someone about to fall down a well nicely illustrates this sentimentalist trend or direction. But unlike Western sentimentalists, the Confucians never directly engage in metaethical questions about the meaning or semantics of moral terms or judgments, and this, together with the above-mentioned fact that they never clearly distinguish reason itself from the sentiments, makes it difficult to say that Mencius was definitely a sentimentalist and Confucius definitely a rationalist. However, the fact that Mencius, like Hume but unlike Confucius and Aristotle, holds that human nature is fundamentally good strengthens to some extent the case for considering Mencius closer to Hume and to sentimentalism than Confucius was.

Now Western philosophers have recently taken more and more interest in Confucian thought and in comparisons with Western virtue ethics, and it is to be expected that the two traditions will influence each other increasingly in the future. (This is already beginning to happen.) And to the extent that both sentimentalism and rationalism are in play within contemporary Western virtue ethics, this distinction will probably be increasingly relevant to the interpretation and to the defense

of particular Confucian ideas or traditions within the large tent of Confucianism. But I should not leave the topic of Asian thought and virtue ethics without at least also mentioning Buddhism and its relevance to the issues we have been discussing. When the present Dalai Lama was asked to summarize the Buddhist view of human life, he did so with a single word: kindness. And the traditional Buddhist emphasis on kindness or compassion can clearly be seen as a kind virtue-ethical sentimentalism. (And is not the Christian/Augustinian view that all of ethics and morality resolves itself into the virtue/imperative of loving others as God loves us rather similar to what the Buddhists have said?) So the issue between rationalist and sentimentalist approaches to virtue ethics seems historically quite important and far from being resolved in or for present-day circumstances, and hence I think it would be helpful now if I spoke of these two trends or tendencies seriatim (in the order just indicated), in order to see how either tradition or both traditions can deal with certain ethical and philosophical issues that any viable, plausible, systematic understanding/theory of ethics needs to address.

Problems for Contemporary Virtue Ethics

In concentrating on the problems sentimentalist and rationalist virtue ethics need (in their different ways) to address, I shall be making some simplifying assumptions. I shall assume, for one thing, that Nietzschean virtue ethics and other more clearly egoistic forms of virtue ethics do not have much chance of being taken seriously in present-day circumstances. (This rules out Epicureanism and Stoicism too, though some will complain that, in the latter case, I am doing this too quickly.) Also, almost every contemporary rationalist virtue ethicist seems to prefer Aristotle to Plato, so I am going to focus, in what follows, on present-day Aristotelianism, its promise and its plausibility. Similarly, no one today seems to prefer the sentimentalist virtue ethics of James Martineau over that of David Hume (though I think there is a great deal that is of value in Martineau's work); so I will focus on Humean-type virtue ethics when I come to consider the prospects and problems of contemporary sentimentalist virtue ethics. And we will begin with Aristotelian-type rationalism.

In recent decades and continuing to this day, Aristotelian views have been more influential than any others within the field or school of virtue ethics. The revival of Humean virtue ethics is a somewhat more recent phenomenon, and though that tendency or trend is now gathering some steam, it is still a minority voice within current virtue ethics (which is now so well established that one need not speak of it any more as reviving: it already *has* revived). But Aristotelianism got a helping impetus from the original article by Anscombe, and having now seen more than one generation of contemporary Aristotelian or neo-Aristotelian virtue ethics, it is appropriate for us to take stock and consider how much it has accomplished

and, perhaps more significantly, how much more it needs to (be able to) accomplish. This will largely consist of considering philosophical problems that have been or can be raised about or against neo-Aristotelian rationalist approaches to virtue ethics.

Neo-Aristotelianism has to some extent shuffled off (as inessential or unhelpful) certain doctrines or tendencies of Aristotle's own views. Aristotle believed that all virtue is one, that one cannot have one of the virtues without having them all, but many neo-Aristotelian virtue ethicists deny or question this (initially unintuitive) "doctrine of the unity of the virtues." More importantly, current-day Aristotelians tend not to subscribe to the Aristotelian "doctrine of the mean," according to which all ethical virtue is a matter of acting or being disposed to act in a "medial" way, in a mean between extremes, as when the courageous individual chooses a path that is somewhere between cowardice and foolhardiness (but closer, in fact, to foolhardiness). Aristotle was "the philosopher" for the late medieval Catholic Church, and Aristotelian virtue ethics was the predominant form of ethical theory during the earlier parts of the modern era. But Aristotelianism got itself into trouble as a result of the doctrine of the mean, when seventeenth-century Dutch jurists argued that certain of our obligations, and certain human virtues, could not reasonably be regarded as lying in mean between extremes the way courage does. If I promise to do something, I either keep my promise or I do not, and the virtue of fidelity to one's promises seems therefore to involve the rejection of just one morally unacceptable alternative, rather than a steering between two unacceptable courses of action. And it is clear too that the importance of such examples and their felt force against Aristotelian views are to some extent a function of the needs and opportunities of modern life. In the Middle Ages, promises could be enforced by communal sanctions based on communal feeling, but in the conditions of modern life, as it was emerging in the (run-up to the) seventeenth century, it is necessary to be able to make promises to and enter into contracts with strangers, and I believe the importance of this factor in modern life brought moral issues about contract and promising increasingly into theoretical consideration – with the result that the Aristotelian model of virtue ethics, with its doctrine of the mean, came increasingly under fire.

However, almost no contemporary neo-Aristotelian subscribes to that doctrine, and there is a further difference between Aristotle himself and present-day Aristotelian virtue ethics that also needs to be considered. Aristotle did not give kindness and compassion a (central) place in his ethical theory, but all the contemporary (neo-)Aristotelians I know of do think of these virtues as morally important. That fact may well reflect the influence of Judeo-Christianity, with its emphasis on compassion and (nonerotic) love, but when such virtues are stressed, the purely rationalistic character of traditional Aristotelianism becomes somewhat attenuated. The Aristotelian who emphasizes things like kindness seems to be taking a page from the sentimentalist book, and if Aristotelianism needs to move in this direction, the case for sentimentalism will seem somewhat strengthened (unless it turns

out that sentimentalism needs to borrow elements from rationalistic Aristotelian virtue ethics).

What actually seems most problematic and/or objectionable about present-day Aristotelianism is not what it has to say about the morality or ethics of individual action, but what it says or has said or needs to say about issues of political morality. Aristotle himself, like Plato, had a rather jaundiced view of the value and validity of democracy, and his official political doctrine favors an aristocracy of virtue over other forms of government. This represents a problem, in contemporary Western circumstances, for those of us who think or feel that only democratic societies can be really or fully just (or morally acceptable). Both utilitarian consequentialism and Kantian ethics in the form of Rawlsian liberalism offer defenses of democratic institutions and values, and the historic failure of Aristotelianism (or any other kind of virtue ethics) to do anything similar therefore counts as a strike against the Aristotelian approach. And it is far from clear how or whether present-day neo-Aristotelianism can sufficiently reverse the historic antidemocratism of Aristotelian ethics. Martha Nussbaum (1990) has pointed out that Aristotle favored a kind of social democracy for and within the class of full citizens, but since this does not tell us how, and with what Aristotelian arguments, slaves and women and farmers are to be brought into the favored circle of politically considerable individuals, it is not clear from her description or Aristotle's how Aristotelian ideas can lead, without distortion or exaggeration, toward some sort of democratic conception of social/political justice.

Rosalind Hursthouse (1991) has also addressed this issue, but though her arguments show how the neo-Aristotelian might be able to justify rights of free speech, assembly, and worship, she says nothing to indicate how full democratic rights – in particular, the right to vote – can be assured by means of the considerations she discusses and emphasizes. And this seems a weakness, because we would, most of us, like to see ethical/political theory provide a justificatory basis for full democratic rights. But there is another serious problem with Aristotelian views (one that is somewhat related to the issue of democratic values).

When Aristotelianism went into decline toward the end of the seventeenth century, it was partly, as I have said, because it offered no plausible way to account for the moral nature and force of promises and contracts. But there was another reason for the decline that may have been even more important than the one just mentioned. The seventeenth century saw the increasing emergence of anonymous large-scale social conditions (as contrasted with what happens in smaller communities) that required the existence of laws governing contracts and the like. But those more modern conditions also involved less homogeneous populations than had occurred or been visible previously. Differences of religion and sect become increasingly a part of the circumstances of modern life in particular societies, and in the absence of common cultural/social values, it becomes more and more imperative that there be laws enabling such diverse and disagreeing groups to live together in the same state or society (not to mention the need for international laws governing transnational commercial and other interactions). In such

pluralistic circumstances, there is a need for tolerance and understanding and for a willingness to put aside disagreements and differences in the name of social and individual good.

However, as Jerome Schneewind (1997) has pointed out, Aristotelianism is very unhelpful with regard to these needs and these issues. Aristotle regarded the virtuous individual as knowing ethical truth in a way that others would not, and he did not think the virtuous would be able to argue or persuade the morally benighted out of their ethical bathos. In fact, Aristotle thought the virtuous individual had no reason to be humble with others or admit he (or she) might be mistaken in ethical matters, and this marked absence of all moral humility is tantamount to a kind of intolerance and ethical arrogance toward the opinions of others – *hardly the basis for the kind of mutual tolerance and understanding that the welfare of a pluralistic society depends on*. These facts about Aristotelianism were well understood in the seventeenth century and earlier, and so in the pluralistic and large-scale social circumstances of that period and later, Aristotelianism was seen as an impediment to the political compromises and toleration that peace and prosperity increasingly depended on. What emerged in its place was a moral doctrine that deliberately and self-consciously accommodates the needs or requirements of pluralistic social conditions, a doctrine that had not, in fact, been articulated earlier, but that also was not as much *needed* under earlier more homogeneous social conditions as it was in the increasingly complex and discordant circumstances of modern life.

And that moral doctrine was the doctrine of basic human (political) rights. Those who accepted such ideas could argue that they had to respect the rights of free speech and religious worship of those who disagreed with them with regard to their most basic religious and even moral beliefs. And so the doctrine of human rights seemed much more useful and helpful in modern circumstances than Aristotelianism appeared to be. Since rights are also not a matter of any mean between extremes, the whole Aristotelian edifice came tumbling down, and Aristotelian virtue ethics was in full eclipse until neo-Aristotelianism emerged in the wake of Anscombe's article. But is it clear how the neo-Aristotelian can adjust the original ideas of Aristotelian ethics so that it no longer stands in the way of modern-day accommodations for toleration toward, and understanding of, those who deeply disagree with one. It is not enough for the neo-Aristotelian simply to declare that humility is (surprise!) a virtue after all. For if the virtuous individual is, as Aristotle says, practically wise, why should they think or admit that their ideas may be fallible and that they may have much to learn from others? Neo-Aristotelianism needs to be able to say something convincing and convincingly Aristotelian about the importance of humility, toleration, and nonarrogance, and I am not aware of anything within the Aristotelian tradition that can enable it to accomplish that end. Unless this happens, the worry that neo-Aristotelian virtue ethics – even one shorn of the original Aristotelian doctrines of the mean and of the unity of the virtues – will be unable to function as an ethical doctrine

for and in modern-day circumstances will remain a pressing one, and it is partly for this sort of reason that I prefer a sentimentalist approach to virtue ethics, one that, as we shall see in a moment, permits one to value and emphasize the importance of humility toward others (others' views) in a contemporaneously useful way. So let us now turn to the problems and prospects of contemporary sentimentalist virtue ethics.

I mentioned earlier that sentimentalism and sentimentalist virtue ethics in particular can be suspected of making ethics an entirely subjective or emotive matter, that the most familiar way in which sentimentalism treats or understands moral utterances is to see them as expressions of emotion that do not necessarily say anything objectively true or valid about the actions, attitudes, or motives they purport to evaluate. But some forms of sentimentalist metaethics make moral claims come out as more objective than do others. The so-called ideal observer theory, for example, treats moral claims as descriptions of our general human sympathetic tendencies rather than as mere emotional expressions of such tendencies (or purely first-person descriptions of how the utterer feels about things). Sentimentalist metaethical emotivism treats "This is wrong" as a sheer *expression* of emotional disapproval (whatever that is); a sentimentalist subjectivism says that "This is wrong" simply *describes* the particular utterer's emotional attitude. But the ideal observer theory claims that when we say that some kind of action is wrong, we are saying that human beings generally have a tendency or disposition to have negative feelings about such actions, and such a general sociological statement or conclusion has some claim, surely, to a certain kind of objectivity. But still, and as rationalists and common sense would probably object, the claim that torturing is wrong seems to make an objective claim about certain kinds of actions rather than referring to or describing our (dispassionate) *reactions* to torture, and the ideal observer theory (and what are called response-dependent views) cannot accommodate this initial and, very likely, persistent intuition. So if that is the best sentimentalist metaethics can do, sentimentalist virtue ethics has a problem that rationalism easily succeeds in avoiding. But *is* that the best?

I have recently come to think that it is not. Making use of the work of Saul Kripke (1981), I am inclined to think that moral claims are anchored in our human sympathetic or empathic dispositions, but that they do not in any way describe those dispositions (Slote 2010). This new metaethical approach also allows for a certain degree of "objective prescriptivity" à la Mackie. It holds that we cannot make moral judgments unless we are capable of empathy and aware of the existence of (our own) empathic reactions, and it argues further that our moral judgments at one and the same time describe an objective moral reality that our sympathetic dispositions serve to "light up" for us and also involve our being motivationally disposed toward the kinds of helping or altruistic behavior that are part and parcel of being sympathetic/empathetic toward others. But unless something along these lines can be made to work, sentimentalism does have a problem with moral objectivity that rationalism finds it much easier to deal with convincingly.

However, there is a nest of other problems with sentimentalism that the sentimentalist virtue ethicist needs to be able to deal with. For example, sentimentalism can speak very relevantly to normative issues concerning the morality of helping, of benevolence, but in order to do so, I believe it has to bring in the notion of empathy in a very specific and emphatic way. Many recent psychologists subscribe to an "empathy-altruism hypothesis" according to which altruism depends on our capacity for empathy with others. And a normative sentimentalism that brings in empathy can therefore claim that acts are wrong if and when they reflect or exhibit a deficiency of empathic concern for others (presumably including animals). But Kantians are always arguing that the emotions are an unreliable basis for moral behavior or action, because (they say) our best intentions, our strongest benevolent motives, sometimes flag or fail in ways that undercut the doing of right actions. This can make one wonder whether sentimentalism possesses the resources to explain the motivation that is necessary to acting rightly. However, the recent literature on empathy and altruism speaks at great length of the ways in which our empathic/altruistic tendencies can be enhanced and strengthened, and I think – and have argued (Slote 2010) – that it is by no means obvious that such moral education or moral training would inevitably leave people without the moral resources to generally do the right thing (even when there are obstacles to doing so).

But morality involves issues of justice in addition to issues of altruistic benevolence, and it is far from clear that, or how, sentimentalism can deal with justice at all, much less in a way that accommodates and justifies what we think about modern-day democratic ideals and institutions. However, the sentimentalist is not totally without resources for handling this kind of issue or problem, and I want to say that the main resource here for the sentimentalist is, once again, the concept of empathy. (I think it can also help us with ordinary individual moral deontology, but I will not say more about that here.)

One feature of empathy that present-day psychologists have not emphasized enough is its "bias" in favor of the badly off as compared with those who are merely not wonderfully off. If we have to choose between giving some help to those who are in desperate need and doing somewhat more good for those who are already well-off but simply not (yet) wonderfully off, we tend to prefer to help those who are worse off, badly off, and this tendency would normally be thought of as in keeping with and even required by the demands of justice. Justice is more concerned with those who are truly badly off than with those who are merely not wonderfully off, and in this respect, it is important to note, our thinking about justice traces the contours of our empathic human tendencies. We have much more (or stronger) empathy for someone who is in terrible straits than for someone whose condition is merely mediocre, and if, as I suggested above, the sentimentalist holds that empathy enters into our moral concepts in a fundamental way, that would both explain and justify our belief that justice requires greater attention to and concern for the badly off than for anyone else (other things being equal). So if I am not mistaken, empathy allows sentimentalist virtue ethics to gain a

foothold on one of the most important aspects of the concept of, and our thinking about, social justice. And empathy is relevant to other issues of justice as well.

For example, political officials in Singapore, China, and elsewhere have argued that the Western insistence on democracy and on the right to vote is an artifact of Western biases and traditional attitudes, one that it makes no sense to try to impose on Asian societies where "natural deference" undercuts the need or advisability of full-scale democratic institutions. But does this deference really come naturally to Asians in a way that it does not come naturally to Westerners, so that these basic psychological differences argue for different kinds of institutions for each kind of society. Well, that is what the Asian officials often tell us, but such claims are eerily and disturbingly similar to the kind of thing that used to be said about natural *female* deference. And we know enough nowadays to be suspicious and more than suspicious of claims that women are naturally deferential, when such claims are used as a basis for moral/political conclusions but also more generally. And here is what I take to be a plausible account of what is morally invidious about the claims as applied to women.

If women were deferential in earlier times, that was probably because people rarely listened to what they had to say or paid attention to their self-liberating aspirations. If a girl said she wanted to be a doctor, her parents would be likely to tell her: you do not really want to be a doctor, dear, you would be much happier in a more feminine calling like nursing or simply staying at home with your children. Such a response does not take what the little girl says about her aspirations seriously, and when a little girl's views and desires do not get a real hearing, the little girl can start doubting herself, start doubting whether she really wants or believes what she initially says she wants or believes. This is the path to deference (and possibly a lot of unacknowledged anger). But consciousness raising (which can tap into the anger) helps the woman who has been such a girl see how disrespectfully, how unempathically, her views and aspirations have been treated by her parents or society at large. And it seems to me that one lesson of the women's movement has been that women are not naturally deferential, but rather are treated in a disrespectful fashion that leaves them deferential for lack of a full or fuller confidence in themselves.

I am inclined to apply the same analysis to the argument for the "Asian value" of deference and the inappropriateness of democracy in such non-Western circumstances. I will bet Asians are not naturally deferential any more than women are. And if Asians acquiesce in authoritarian regimes and accept their own deference as a given or inevitable, that may very well be a sign of the same kind of unempathic and disrespectful treatment that has led women to be deferential. If, from childhood on, women and Asians *are* listened to, there will be, I think, no natural deference, and so the typical contemporary argument against democracy that comes to us out of Asia seems to be as flawed as arguments for denying deferential women the vote. And since, once again, empathy with the point of view of others is a key moral feature of our criticism of current arguments against democracy (in the Asian context), we have more evidence of how sentimentalist virtue ethics

that emphasizes empathically based benevolence and altruism can also account for our received views and strong intuitions about social justice.

Finally, let me mention the right to worship freely. That moral/political right is often justified or defended in rationalist terms, but sentimentalism too has something useful and persuasive to say about freedom of religion, and once again this will depend crucially on the concept of empathy. It is sometimes said that we need to have and to invoke a rationally grounded right of free religious worship because positive sentiments like benevolence and compassion could easily lead one to deny people the right to worship as they please: as when those who persecute heretics say that they have to do so for the sake of the eternal welfare of (the souls of) those who hold false heretical beliefs. But as John Locke wisely and wryly pointed out in his *Second Essay on Government*, the "dry eyes" of those who torture people for the supposed good of their souls refute the claim that the torturers really have the welfare of their victims at heart (1960). What most pervasively characterizes religious persecution and acts of torture, rather, is hostility, anger, and contempt toward those who disagree with the religious views of those in power, together, often, with a desire to confiscate and gain the property (and positions) of those who are persecuted and "broken" or killed. Has it really ever been any different?

In that case, it seems plausible to hold that if people were more empathic with and less hostile toward the views of those who disagree with them, there would be no such thing as religious persecution. (Issues of greed can be handled in separate fashion.) And one can therefore, I think, both plausibly and comfortably say that what is, at the most human level, wrong with and unjust about persecuting people for their beliefs is an unwillingness or failure to engage with those differing beliefs, the lack of any effort to seeing things from the standpoint of those others. In other words, a lack of empathy. But what also emerges from this discussion is that a sentimentalist virtue ethics that emphasizes empathy is in a far different position from historical and contemporary Aristotelianism regarding the issues/problems of mutual tolerance and understanding that are so prominent and pressing in modern-day pluralistic societies (or between societies). Unlike Aristotelian rationalism, the idea and phenomenon of empathy are or can be fundamental to virtue-ethical sentimentalism, and therefore it is no stretch, it is very much in keeping with sentimentalism, to speak of the need for and ethical desirability of empathy for other people's views and ideas and thus to invoke (the value of) a kind of humility that Aristotelianism seems to have no room for. So I am saying that sentimentalist virtue ethics is better geared for, more appropriate to, modern or contemporary social/political circumstances than any form of Aristotelianism I know of.

But this is far from settling the issue as between sentimentalist and rationalist virtue ethics. Both sorts of approaches face intellectual/philosophical challenges (often different ones) that they have not satisfactorily responded to, and only time will tell whether either sort of approach will be able to deal more successfully with

moral and political issues than Kantianism or consequentialism will prove capable of doing.

References

Anscombe, G.E.M. (1958) "Modern Moral Philosophy," *Philosophy* 33: 1–19.

Baier, Annette (1985) "What Do Women Want in a Moral Theory?" *Noûs* 19: 53–63.

Blackburn, Simon (1984) *Spreading the Word*, Oxford: Oxford University Press.

Confucius (Kongzi) (1979) *The Analects*, ed. D.C. Lau, New York: Penguin.

Hume, David (1739/1978) *A Treatise of Human Nature*, ed. L.A. Selby-Bigge and P. Nidditch, Oxford: Oxford University Press.

Hursthouse, Rosalind (1991) "After Hume's Justice," *The Aristotelian Society* 91: 229–45.

Hursthouse, Rosalind (1999) *On Virtue Ethics*, Oxford: Oxford University Press.

Kant, Immanuel (1964) *Doctrine of Virtue*, New York: Harper.

Kripke, Saul (1981) *Naming and Necessity*, Oxford: Blackwell.

Kuhn, Thomas (1962) *The Structure of Scientific Revolutions*, Chicago: University of Chicago Press.

Locke, John (1960) "Second Treatise," in *Two Treatises of Government*, ed. Peter Laslett, Cambridge: Cambridge University Press, pp. 265–428.

Mackie, J.L. (1977) *Ethics: Inventing Right and Wrong*, Harmondsworth, UK: Penguin.

Martineau, James (1891) *Types of Ethical Theory*, Oxford: Clarendon Press.

McDowell, John (1979) "Virtue and Reason," *The Monist* 62: 331–50.

Mencius (Mengzi) (1970) *Mencius*, ed. D.C. Lau, New York: Penguin.

Nussbaum, Martha (1990) "Aristotelian Social Democracy," in *Liberalism and the Good*, eds. R.B. Douglass, G. Mara, and H. Richardson, London: Routledge.

Pincoffs, Edmund (1986) *Quandaries and Virtues*, University of Kansas Press.

Rawls, John (1971) *A Theory of Justice*, Cambridge, MA: Harvard University Press.

Schneewind, Jerome (1997) "The Misfortunes of Virtue," in *Virtue Ethics*, eds. R. Crisp and Michael Slote, Oxford: Oxford University Press, pp. 178–200.

Slote, Michael (2010) *Moral Sentimentalism*, New York: Oxford University Press.

Stephen, Leslie (1882) *The Science of Ethics*, New York: G.P. Putnam's Sons.

Stevenson, Charles (1944) *Ethics and Language*, New Haven: Yale University Press.

Williams, Bernard (1985) *Ethics and the Limits of Philosophy*, Cambridge, MA: Harvard University Press.

Chapter 19

Capability Ethics

Ingrid Robeyns

The capability approach is one of the most recent additions to the landscape of normative theories in ethics and political philosophy. Yet in its present stage of development, the capability approach is not a full-blown normative theory, in contrast to utilitarianism, deontological theories, virtue ethics, or pragmatism. As I will argue in this essay, at present the core of the capability approach is an account of value, which together with some other (more minor) normative commitments adds up to a general normative framework that can be further developed in a range of more specific and detailed normative theories. The aim of this essay is to describe the capability approach, as it has been developed so far, as well as briefly explore how a capabilitarian ethical theory could look if we were to develop it in full.

So what is the capability approach? In its most general description, the capability approach is a flexible and multipurpose normative framework, rather than a precise theory of well-being and freedom. At its core are two normative claims: first, that the freedom to achieve well-being is of central moral importance, and second, that the freedom to achieve well-being is to be understood in terms of people's valuable capabilities, that is, their real opportunities to do and be what they have reason to value. This framework can be used for a range of evaluative exercises, including most prominently the following: (1) the assessment of individual well-being; (2) the evaluation and assessment of social arrangements, including assessments of social and distributive justice; and (3) the design of policies and proposals about social change in society, which is at the core of social ethics. In all these normative endeavors, the capability approach prioritizes (a selection of) people's beings and doings and their opportunities to realize those beings and doings – for example, their genuine opportunities to be educated, their ability to move around or to enjoy supportive social relationships. This stands in contrast to normative

The Blackwell Guide to Ethical Theory, Second Edition. Edited by Hugh LaFollette and Ingmar Persson.

frameworks which endorse other accounts of value, like mental states, or which focus on instrumental values, like resources.

What Are Capabilities?

Capabilities are real opportunities, whereby the opportunities do not refer to access to resources or to certain levels of satisfaction but rather to what a person can do and to the various states of being of this person. Capabilities refer to both what we are able to do (activities), as well as the kind of person we can be (dimensions of our being). These "beings and doings" – that is, the various states of human beings and activities that a person can undertake – are in the capability approach called "(human) *functionings.*" The term "functionings" may misleadingly sound like a profound notion, but in essence it refers to a very simple idea that we use in common speech, namely aspects or states of the human being that we are ("beings"), and the various things we can do as human beings that we are ("doings"). Examples of the former (the "beings") are being well nourished, being undernourished, being housed in a pleasantly warm but not excessively hot house, being educated, being illiterate, being part of a supportive social network, being part of a criminal network, and being depressed. Examples of the second group of functionings (the "doings") are traveling, caring for a child, voting in an election, taking part in a debate, taking drugs, killing animals, eating animals, consuming lots of fuel in order to heat one's house, and donating money to charity.

Although many scholars who are working with the capabilities approach focus exclusively on capabilities that they deem valuable, conceptually the notion of "functionings" is in itself morally neutral. Functionings can be univocally good (e.g., being in good health) or univocally bad (e.g., being raped). But the goodness or badness of various other functionings may not be so straightforward, but rather depend on the context and/or the normative theory that we endorse. For example, is the care work of a mother who is caring full time for her child a valuable functioning or not? A conservative-communitarian normative theory will most likely mark this as a valuable functioning, whereas a liberal-feminist theory will only do so if the care work is the result of an autonomous choice made against a background of equal opportunities and fair support for those who have duties to care for dependents.

Capabilities are a person's real freedoms or opportunities to achieve functionings. Thus, while traveling is a functioning, the real opportunity to travel is the corresponding capability. A person who is not traveling may or may not be free and able to travel: the notion of "capability" tries to capture precisely this fact of whether the person *could* travel if she would want to do this. The distinction between functionings and capabilities is between the realized and the effectively possible, in other words, between achievements, on the one hand, and freedoms or valuable opportunities from which one can choose, on the other.

The notion of "capabilities" refers to a combination of external conditions of acting or being combined with internal conditions. Take for example the capability to read. The internal conditions for that capability include having decent reading skills, as well as not suffering from debilitating conditions that prevent a person from reading (e.g., severe concentration issues or visual health problems). The external conditions for that capability are, for example, having access to a text that is written in a language one masters, as well as being in the right environment and conditions which allow one the time and space to read (e.g., not being in a situation of acute physical danger).

Martha Nussbaum acknowledges the importance of both the internal and external conditions explicitly, by working out the conceptual distinction between internal capabilities and combined capabilities. Internal capabilities are "the characteristics of a person (personality traits, intellectual and emotional capacities, states of bodily fitness and health, internalized learning, skills of perception and movement)" (Nussbaum 2011a: 21). Combined capabilities, which in the capability literature are referred to simply as "capabilities," are "not just the abilities residing inside a person but also the freedoms or opportunities created by a combination of personal abilities and the political, social and economic environment" (Nussbaum 2011a: 20). Combined capabilities, or "capabilities" for short, are thus the internal capabilities in combination with the relevant external conditions.

Capabilities are often referred to as "freedoms." Yet this can only coherently be done if freedoms are understood in a very particular sense, since the term "freedom" is used (both in daily speech and in practical philosophy) in a range of different ways, many of which are not what is meant by those who equate capabilities with freedoms. This is one reason why Nussbaum (2003) is critical of calling capabilities freedoms, since it is a language that is often equated with the rule of law, formal rights, and similar formal or legal notions. She is especially worried that in certain contexts the reference to "freedoms" will lead to an understanding of capabilities as implying mere noninterference, whereas the protection or enhancement of a person's capabilities often requires proactive policies. Nussbaum has on a few occasions used the term "substantial freedoms" (e.g., 2011a: 21) but generally has preferred the use of the term "opportunities" when describing capabilities.

Amartya Sen, in contrast, often equates capabilities with freedoms. Yet as any overview on the concept of "freedom" will quickly reveal, "freedom" has many different and often conflicting meanings (see, for example, Carter, Kramer, and Steiner 2007). A careful reading of Sen's work clarifies that capabilities are freedoms conceived as real opportunities (Sen 1985a: 3–4; 1985b: 201; 2002: ch. 20). For Sen, capabilities as freedoms refer to the *presence* of valuable options or alternatives, in the sense of opportunities that do not exist only formally or legally but are also effectively available to the agent. As Alexander Kaufman (2006) has shown, some critiques of Sen's writings on the capability approach are based on a mistaken understanding of the kind of freedom capabilities are.

Many ethical theories require a metric for interpersonal comparisons. According to the capability approach, "functionings" and "capabilities" are the best metric

for most kinds of interpersonal evaluations. Yet the various versions of the capability approach differ in whether they have an underlying theory of reasons to value (and if so, which), and whether capabilities are strictly tied to people's well-being. In general, there is very little explicit discussion about these issues in the capability approach, which makes situating this theory in the landscape of ethical theories tricky. In fact, one could argue that nothing can be concluded about the capability approach in general, but that rather each particular capability theory needs to analyzed on its own before its theoretical properties can be properly described.

In Sen's version of the capability approach, interpersonal evaluations should be conceptualized in terms of people's capabilities to function, that is, their effective opportunities to undertake actions and activities that they have reason to value, and be the person that they have reason to want to be. These beings and doings together are held to constitute what makes a life valuable. Whereas "functionings" are the proposed conceptualization for interpersonal comparisons of (achieved) well-being, "capabilities" are the conceptualization for interpersonal comparisons of the freedom to pursue well-being, which Sen calls "well-being freedom" (Sen 1992: 40). Sen's claim is that functionings are constitutive of a person's being, and an evaluation of well-being has to take the form of an assessment of these constituent elements (Sen 1992: 39).

Sen's account of the capability approach is deliberately vague and ambiguous – in part because he believes the notion of capabilities can be helpful for different normative goals. Two points should be noted here. First, Sen has a broad and plural approach to ethical evaluation: he holds that there is not a single value which is always more important than other values. In particular, he argues that we can make a distinction between achieved well-being (reflected in a person's functionings), well-being freedom (reflected in her capabilities), as well as achieved agency and agency freedom. Both aspects of agency are for Sen very important, and evaluating agency requires the evaluation to go beyond functionings and capabilities. As a consequence, Sen holds that a person's functionings and capabilities are not all that matters: she may have certain agency goals, which translate into particular commitments which may harm her own well-being. For example, an environmental activist or a human rights advocate may put at risk their own well-being, yet for Sen there may be good reasons why an ethical evaluation should also assess whether people can pursue their agency goals (Sen 1985b).

The second particular feature of Sen's capability theory is the role played by "reasons to value" in his account. In Sen's view, we need a process of public reasoning in order to find out which capabilities are valuable. Sen (2009b: 31–51) believes that there cannot be a guarantee that by reasoning we will reach the truth, but that this should not prevent us from trying to be as objective as we can. While in the context of the capability approach Sen has said little on what the theory of reasons is that underlies his claim that the capability approach should focus on the capabilities which people have *reason to value*, he has indicated repeatedly that he believes this should be a process of *public* reasoning. He draws on Adam Smith's notion of the impartial spectator in order to bring in the necessary objectivity in

public reasoning, since Smith's method requires us to broaden the discussion about reasons beyond the local perspectives and interests. Sen (2009b: 45) does indeed "take reasoned scrutiny from different perspectives to be an essential part of the demands for ethical and political convictions." While a critic may remark that Sen's writings on the role of "reasons to value" in his capabilities theory remain vague and perhaps even ambiguous, a defender of Sen could stress that it is a mistake to believe that one can do more than outline a procedure, as there is no final account of such reasons to be given.

Nussbaum's version of the capability approach has very different normative foundations. There is no role for public reasoning in her theory, apart from the level of implementing the capability approach in a particular context where local people need to be involved in translating the abstract capabilities into more specific capabilities that are sensitive to the local particularities. Nussbaum has been very clear and explicit for a long time that her capabilities approach is a partial theory of social justice: it is defending an account of a minimal set of capabilities which all people on earth should be entitled to purely on grounds of their humanity. Over time, Nussbaum has developed and defended a list of "central human capabilities," entailing the following ten broad categories: life; bodily health; bodily integrity; senses, imagination and thought; emotions; practical reason; affiliation; other species; play; and control over one's environment (Nussbaum 2006: 76–8). She believes these ten categories of capabilities would be endorsed by people across cultures and circumstances. In other words, Nussbaum claims that the ten categories of capabilities on her list of central human capabilities form the object of a Rawlsian overlapping consensus: independent of people's own comprehensive views of the good life, all will agree that as members of a common humanity, we owe each other minimal levels of these ten capabilities. This would turn Nussbaum's capabilities approach into a form of political liberalism, which she believes is the only defensible form of liberalism in political philosophy (Nussbaum 2011b). Yet as is the case with Sen too, Nussbaum's work attracts many critics, including many who oppose her *way* of doing philosophy, or who argue that her version of the capabilities approach is a form of liberal perfectionism (e.g., Barclay 2003).

While Sen and Nussbaum ultimately develop very different versions of the capability approach, which are based on quite different ethical foundations and are ultimately based on conflicting views of what a theory can and ought to do, they do of course also share some common theoretical ground. One important aspect which all capability theories share is the idea that functionings are constitutive for human beings. To say that functionings are constitutive of a person's being means that one cannot be a human being without having at least a range of functionings: they make the lives of human beings both lives (in contrast to the existence of innate objects) and also human (in contrast to the lives of trees or termites). Human functionings are those beings and doings that we take to constitute a human life, and which are central in our understandings of ourselves as human beings. This implies that the range of potentially relevant functionings is very broad, and that the capability approach will in some respects be close to both

subjective metrics (for example, by including the capability to be happy), or resources-based metrics (since most functionings require some resources as inputs). Yet not all beings and doings are functionings; for example, being able to fly like a bird or reaching an age of 200 like an oak tree are not human functionings.

Thus, according to the capability approach, the ends of well-being, freedom, and development should be conceptualized in terms of people's capabilities. Moreover, what is relevant is not only which opportunities are open to me each by themselves, hence in a piecemeal way, but rather which overall real opportunities I have in life – in technical terms: which combinations or sets of potential functionings are available to me. For example, suppose I am a low-skilled poor single parent who lives in a society without decent social provisions. Take the following functionings: (1) to hold a job, which will require me to spend many hours working and commuting, but will generate the income needed to properly feed myself and my family; (2) to care for my children at home and give them all the attention, care, and supervision they need. In a piecemeal analysis, both (1) and (2) are opportunities open to me, but they are not both together open to me. The point about the capability approach is precisely that we must take a comprehensive or holistic approach, and ask which sets of capabilities are open to me; that is, can I simultaneously provide for my family and properly care for my children? Or am I rather forced to make some hard, perhaps even tragic choices between two functionings which both reflect basic needs and basic moral duties?

Note that while most types of capability analysis require interpersonal comparisons, one could also use the capability approach to evaluate the well-being or well-being freedom of one person at one point in time (e.g., evaluate his situation against a capability-yardstick), or to evaluate the changes in his well-being or well-being freedom over time. The capability approach could thus also be used by a single individual in his deliberate decision making or evaluation processes, but these types of uses of the capability approach are much less prevalent in the philosophical literature, let alone in the social sciences.

In development ethics, or in ethical theories that propose to deal with poverty issues, the notion of "basic capabilities" may play an important role too. The term "basic capabilities" refers to a threshold level for the relevant capabilities. A basic capability is "the ability to satisfy certain elementary and crucially important functionings up to certain levels" (Sen 1992: 45 n. 19). Basic capabilities refer to the freedom to do some basic things considered necessary for survival and to avoid or escape poverty or other serious deprivations. The relevance of basic capabilities is "not so much in ranking living standards, but in deciding on a cut-off point for the purpose of assessing poverty and deprivation" (Sen 1987: 109). However, while this is the most widespread use in the present development of the capability approach, Nussbaum (2000: 84) uses the term "basic capabilities" to refer to "the innate equipment of individuals that is necessary for developing the more advanced capabilities," such as the capability of speech and language, which is present in a newborn but needs to be fostered. The lesson to take home for the student of the capability approach who delves into the primary texts is that not all

capability scholars are using the exact same terminology. There are differences between the terms used by Sen and Nussbaum, but there are also differences between the various disciplines and subdisciplines in which the capability approach is being developed.

The Importance of Human Diversity

A deep acknowledgement of human diversity is one of the key theoretical characteristics of the capability approach. Its criticism of other normative approaches is often fueled by, and based on, the claim that the full human diversity among people is insufficiently acknowledged in many normative theories, such as theories of distributive justice. This also explains why the capability approach is often favorably regarded by feminist philosophers, or philosophers concerned with care and disability issues, since one of their main complaints about mainstream moral philosophy has precisely been the relative invisibility of the fate of those people whose lives do not correspond to that of an able-bodied, nondependent, care-duties-free individual who belongs to the dominant ethnic, racial, and religious group. People of minority ethnicities, marginalized people, the disabled, and many women do not fit that picture.

Where does this concern for human diversity manifest itself in the structure of the capability approach? The first and most obvious aspect of the capability approach that reveals its concern for human diversity is that the approach in general focuses on all functionings that are relevant and of value – and that generally tends to be a wide range of dimensions. Compared with normative theories that include a metric of value, it tends to focus on a broader range of goods, thereby often including dimensions that are particularly important for some groups, but not for others. Hence, when thinking for example about the good life and a decent or just society, issues of care and dependency will be central to the ethical assessment, which is often not the case in mainstream ethics or normative political philosophy.

Secondly, human diversity plays a crucial role in the reason why the capability approach focuses on capabilities and functionings rather than resources. The capability approach argues that our evaluations and acts should start from what has ultimate value (which it holds to be capabilities), and only in a second step of the analysis ask what means (broadly defined) are needed to secure these capabilities. This implies that the capability approach evaluates policies and other changes according to their impact on people's capabilities as well as their actual functionings. It asks whether people are able to be healthy, and whether the means or resources necessary for this capability, such as clean water, adequate sanitation, access to doctors, protection from infections and diseases, and basic knowledge on health issues, are present. It asks whether people are well nourished, and

whether the means or conditions for the realization of this capability, such as having sufficient food supplies and food entitlements, are being met. It asks whether people have access to a high-quality education system, to real political participation, and to community activities that support them, that enable them to cope with struggles in daily life, and that foster caring and warm friendships.

Apart from the fact that capability scholars (in line with most moral philosophers but not with the practice of many economists and policymakers) believe that one should start any ethical analysis from what really matters (rather than from the means to those goods), human diversity is the main reason to focus on ends (capabilities) rather than means. The reason is that people differ dramatically in their ability to convert resources into valuable functionings. That implies that, as a matter of public policy and social ethics, creating capabilities will require different actions for different groups of people. People are unique in their personal characteristics, and in how they fare in the communities and environment in which they are embedded and located. Some of these factors which create interindividual diversity are body related, others are shared with all people from her community, and still others are shared with people with the same social characteristics (e.g., same gender or class or race characteristics). If we want to give a child with autism maximal chances to achieve a valuable set of capabilities, then he will need a teacher with different skills and knowledge, and a different school setting (e.g., smaller class size) when compared with a neurotypical child. Similarly, a man wanting to work as a childminder may, given existing gender norms which discriminate against men working with infants, need more social support in order to be able to do the work he wants to do than a woman. Simply giving the child with autism access to public schooling made for neurotypical children, or the male childminder access to a gendered labor market, will not be enough in order to secure that all have the same capabilities to access. Capability scholars believe that these interindividual differences are far-reaching and significant, and that normative theories that focus on means tend to downplay their normative relevance (such theories are perhaps particularly dominant in political ethics or normative political philosophy, and perhaps less in other parts of ethics).

Yet one could wonder to what extent the capability approach does make a relevant difference for ethical analysis, compared to equality of opportunity theories, or to theories that focus on means. I will briefly look at both alternatives in turn.

First, would not it be better to focus on means only, rather than making the normative analysis more complicated and more informationally demanding by also focusing on functionings and capabilities (e.g., Pogge 2002)? Capability scholars would respond that starting a normative analysis from the ends rather than means has at least two advantages, apart from the earlier mentioned fundamental reason that a focus on ends is needed to appropriately capture interindividual differences. First, the valuation of means will retain the status of an instrumental valuation rather than take on the nature of an intrinsic valuation. For example, money or

economic growth will not be valued for their own sake, but only in so far as they contribute to an expansion of people's capabilities. Second, by starting from ends, we do not a priori assume that there is only one overridingly important means to that ends (such as income), but rather explicitly ask the question which types of means are important for the fostering and nurturing of a particular capability, or set of capabilities. For some capabilities, the most important means will indeed be financial resources and economic production, but for others it may be particular political practices and institutions, such as effective guarantees and protections of freedom of thought, political participation, social or cultural practices, social structures, social institutions, public goods, social norms, and traditions and habits. As a consequence, an effective capability-enhancing policy may not exist in increasing disposable income, but rather fighting a homophobic, ethnophobic, racist, or sexist social climate.

A second alternative framework, which could also claim to pay a lot of attention to human diversity, comprises sophisticated equality of opportunity theories. Ronald Dworkin (2002: 299–303) has argued that the capabilities approach, when disambiguated, collapses into something very similar to an equality of opportunity theory. Similarly, the differences with John Rawls's principles of justice have also been analyzed. More detailed analysis has shown that the differences between the specific normative principles defended by these theories are small, both when analyzed in general terms and for particular problems such as health, disabilities education, or gender issues (Brighouse and Robeyns 2010; Pierik and Robeyns 2007; Williams 2002).

Yet there are significant differences at the metatheoretical level, since the capability approach is less committed to a number of specifications that are needed to turn a broad and arguably somewhat vague normative commitment idea into a well-defined theory: its only firm normative commitment is to functionings and capabilities as the relevant evaluative space. Equality of opportunity theories either have certain background assumptions that turn them into ideals that hold only under idealizing circumstances or they have a much more restricted theoretical scope, such as being limited to people in their capacity as citizens only, or only to people who have the cognitive capacities needed for practical reasoning. In addition, there are also differences in terms of the grounding of those theories, and to the basic ethical intuitions from which they start. In sum, the capability approach can properly be regarded as one version of an opportunity theory (after all, capabilities are opportunity freedoms), but it remains a more open framework with fewer specifications, and with, *in general*, no commitment to restrictions on scope (yet more specific capabilitarian theories may include such a commitment). Moreover, the capability approach stresses the important aspect of analyzing *all* relevant areas of life, hence endorsing what some would prefer to call a holistic viewpoint. This implies means that it would reject an analysis that only looks at opportunities in one sphere of life, but would rather ask how a certain act or social institution affects the quality of people's lives and their freedom to have a high quality of life in all relevant domains.

Specifying the Capability Approach

The capability approach defends a specific view of ultimate value by conceptualizing a metric of well-being (in terms of functionings) and well-being freedom (in terms of capabilities). However, clearly this still leaves open a range of very different capability theories to emerge from these metrics. These theories will differ regarding *their purpose* (e.g., a capabilitarian theory of ethics, or a capabilitarian theory of global justice) but they will also differ regarding their *specifications*, which are needed in case we want a theory that gives precise answers to the questions of what actions we ought to take.

In the capability literature, it is generally accepted that at least three specifications are needed. First, is the appropriate focus functionings, or rather capabilities? Second, how are we to select and aggregate the multiple dimensions of the capability approach? And finally, since the capability approach only defends the informational space in which interpersonal comparisons need to be conducted, what else is needed for a full capability theory of justice to be developed? In what follows each will be discussed briefly; a more elaborate discussion, including considerations that are of less relevance for ethical theory but more to the capability approach in the social sciences, can be found in Robeyns (2011).

Functionings or Capabilities?

Scholars interested in the capability approach have debated whether the appropriate well-being metric should be capabilities or functionings, hence opportunities or achievements. What considerations have been argued to be relevant for this choice?

The first consideration is normative, and this is the argument Sen and Nussbaum most often offer: by focusing on capabilities rather than functionings, we do not privilege a particular account of good lives but instead aim at a range of possible ways of life from which each person can choose. Thus, it is the liberal nature of the capability approach, or an antipaternalist consideration, that motivates a principled choice for capabilities rather than functionings. Obviously, the strength of this argument depends on how bad one takes paternalism to be. There may be good reasons to believe that some paternalism is unavoidable, or even desired (Nussbaum 2000: 51–6). Moreover, many would hold that there is most likely some paternalism in the selection of capabilities anyway.

A second normative consideration stems from the importance given to personal responsibility in contemporary political philosophy. If one believes that one should strive for equality of capability (as some but not all capability scholars do since some would rather defend a sufficientarian capability view, which holds that threshold levels of capabilities need to be met for all people, but beyond those thresholds inequalities are morally permissible), then each person should have the same real

opportunity (set of capabilities), but once that is in place, each individual should be held responsible for his or her own choices. This responsibility-sensitivity principle is widely endorsed not only in political philosophy but also in the mathematical models being developed in normative welfare economics. If one wants to endorse and implement this principle of responsibility-sensitivity, then specifications and applications of the capability approach should focus on capabilities, rather than functionings. Yet many philosophers disagree on whether we should endorse responsibility-sensitivity in developing the capability approach (e.g., Fleurbaey 2002; Vallentyne 2005; Wolff and de-Shalit 2007). Moreover, for applied ethical analysis, serious epistemological hurdles may ultimately lead us to drop the responsibility-sensitive principle for practical reasoning about the actual world.

Third, there are cases in which a capability is available to a person but only if other people do not also want to realize that capability (Basu 1987: 74). For example, two spouses may each have the capability of holding demanding jobs which are each on their own incompatible with large caring responsibilities. However, if these spouses also have infants or relatives with extensive care needs, then at best only one of them may effectively realize that capability. Since capability sets may therefore include freedoms that are conditional (because they depend on the choices of other people), it might be better to focus both on the individual's capability set and also on what people have been able to realize from their own capability sets, that is, their functionings or well-being achievements. The question of who decides or should decide this sort of spousal question highlights the importance of agency and procedural fairness, which are generally taken to be part of the capability approach in its broader use (Crocker 2008).

It should also be mentioned that the concept of functioning has particular relevance for our relations to those human beings who are not yet able to choose (infants), who will never be able to choose (severely mentally disabled individuals), or who have lost this ability through advanced dementia or serious brain damage. Whether or not these persons can decide to be well nourished and healthy, it is generally held that we (through families, governments, or other institutions) have the moral obligation to promote or protect their nutritional and healthy functioning.

Selecting and Aggregating of Capabilities?

Other major points of debate in the capability literature are the questions of which capabilities should be selected as relevant and who should decide (or how a decision should be made) on the aggregation of the various dimensions into an overall assessment. At the level of ideal theories of justice, some have argued that each and every capability is relevant and should count in our moral calculus (Vallentyne 2005). Others have argued that considerations of justice require that we demarcate

morally relevant from morally irrelevant and morally bad capabilities (Nussbaum 2003; Pogge 2002; Pierik and Robeyns 2007). This demarcation could be done in various ways, and most capability scholars think that different answers are appropriate in different normative exercises. In other words, the selection of relevant capabilities would be different when the question is how to arrange a society's basic structure, versus when the question is how to spend the donations Oxfam has collected, or when the normative question is how to raise one's child. Anderson (1999) argues that, for purposes of political justice, the only relevant capabilities are those needed for a person to participate as a citizen. Nussbaum endorses a well-defined list of capabilities, which, she argues, should be enshrined in every country's constitution (Nussbaum 2000, 2003, 2006). Her list contains the earlier mentioned ten central human capabilities: life; bodily health; bodily integrity; senses, imagination and thought; emotions; practical reason; affiliation; other species; play; and control over one's environment. Nussbaum (2000: 70–7; 2006: 78–81) justifies this list by arguing that each of these capabilities is needed in order for a human life to be "not so impoverished that it is not worthy of the dignity of a human being" (2000: 72). She defends these capabilities as being the moral entitlements of every human being on earth. She formulates the list at an abstract level and advocates that the translation to implementation and policies should be done at a local level, taking into account local differences. However, Nussbaum is crystal clear that her project is only the formulation of a partial theory of political justice, and hence it is not obvious at all that the same list can be used for other normative projects, nor that the justification she offers for her list can be transposed.

Sen consistently and explicitly refuses to defend "one pre-determined canonical list of capabilities, chosen by theorists without any general social discussion or public reasoning" (Sen 2005: 158). Of course, groups and theorists might construct lists for various purposes, and lists need not be "pre-determined" or "canonical," however we might understand these terms. And Sen's refusal to endorse Nussbaum's list has not prevented him from using – for various purposes – particular selections of capabilities in his empirical as well as his normative work. However, beyond stating in general terms that some democratic process and public reasoning should be involved, Sen has never explained in detail how such a selection could and should be done. Several capability scholars, including Anderson, Alkire, Robeyns, and Crocker, have sought in various ways to fill this lacuna. Anderson (1999: 316) argues that people should be entitled "to whatever capabilities are necessary to enable them to avoid or escape entanglement in oppressive social relationships" and "to the capabilities necessary for functioning as an equal citizen in a democratic state." Alkire (2002: ch. 2) proposes to select capabilities based on John Finnis's practical reasoning approach. By iteratively asking "Why do I do what I do?" one comes to the most basic reasons for acting: life, knowledge, play, aesthetic experience, sociability (friendship), practical reasonableness, and religion. Robeyns (2003) has proposed some pragmatic criteria, mainly

relevant for empirical research, for the selection of capabilities for the context of inequality and well-being assessments. Crocker (2008: chs. 9–10) explores the theory and practice of deliberative democracy to bring more specificity to democratic procedures and participatory institutions in the development of an agency-sensitive capability approach.

What about weighting different capabilities to come to an aggregate evaluation? If we have a list of relevant capabilities, we would still be left wondering whether the capabilities should be aggregated and, if so, what their relative weights and the formula to aggregate them will or should be. A closely related question is how different capabilities should be traded off against one another when they cannot all be realized fully. Some have argued against trade-offs on the basis that the different capabilities are incommensurable or that each capability is an absolute entitlement that never should be overridden by another entitlement or other normative consideration. For example, Nussbaum argues that the ten capabilities on her list, being incommensurable, cannot be traded off against one another (and, hence, have no relative weights), and also that the state should provide each citizen with a minimum threshold of each capability.

One possible system of weighting or aggregating is to use a democratic or some other social choice procedure (Chakraborty 1996). The basic idea would be to encourage or prescribe that the relevant group of people decide on the weights. In some contexts, such as small-scale projects or evaluations, such capability weighting (and selection) could be done by participatory techniques. It has also been suggested that we may determine the weights of capabilities as a function of how much they contribute to overall life satisfaction or happiness (Schokkaert 2007). Yet this raises the question to what extent functionings are taken to be merely instrumental to another end, such as happiness, or indeed any other ultimate good or ideal.

Much of the existing literature refers to the issue of "weighting," but this is only one particular form of the more general "aggregating," since aggregation may take a different functional form than simply adding up. For example, if you have no food, your other capabilities will be worth very little. Some capabilities may thus be complementary capabilities, implying that their value to a person depends on the presence (or absence) of others. (Note the similarity with the notion of "complementary goods" in consumer theory in economics, where it is argued that the utility of some goods is dependent on the quantity of some other goods, as in the case of pencils and erasers, or shoe polish and shoes of the same color.)

Towards a More Complete Capability Theory

The capability approach is often wrongly taken to be an egalitarian theory or a theory of social or distributive justice. This reading is mistaken, even though it is entirely understandable given the specific debates in which the main philosophers

defend this approach. The capability approach specifies what should count for interpersonal evaluations and thus provides an important aspect of a theory of social or distributive justice, or a normative ethical theory, yet more is needed. This implies that the capability approach *can be* an egalitarian theory, but *can also be* a sufficientarian theory; at the highly general level at which the capability approach is pitched, all these more specific theories can be developed.

Within the capability approach, most work has been done in political philosophy, on developing the capability approach into a more detailed theory of social justice. Nussbaum's work comes closest to offering us a capability theory of justice, but her theory too does not amount to a full theory of social justice. Nussbaum's theory of social justice is comprehensive, in the sense that it is not limited to an account of political justice, or to liberal democracies. Rather, her account holds for all human beings on earth, independently of whether they are living in a liberal democratic regime, or of whether they are severely disabled. The main demarcation of Nussbaum's account is that it provides only "a partial and minimal account of social justice" (Nussbaum 2006: 71) by specifying thresholds of a list of capabilities that governments in all nations should guarantee to their citizens. Nussbaum's theory focuses on thresholds, but this does not imply that reaching these thresholds is all that matters for social justice; rather, her theory is partial and simply leaves unaddressed the question of what social justice requires once those thresholds are met.

Moreover, it would be a mistake to think that there can be only one capability theory of justice; on the contrary, the open nature of the capability approach allows for the development of a family of capability theories of justice. But this prompts the questions: What is needed to develop a full capability theory of justice, and which of these aspects have already been developed by theorists of justice? And what is needed for a convincing and plausible capabilitarian ethical theory? Since on that last question very little work has been done so far, I will develop some first thoughts in the last section, "Towards an Ethical Theory," of this essay. On the question of what is needed for a plausible and complete capability theory of justice to emerge, see Robeyns (2011).

Capabilities and Rights

The notion of rights plays a central role in practical philosophy, but also in our daily lives. As Nussbaum (1997: 273) has pointed out, "The language of rights has a moral resonance that makes it hard to avoid in contemporary discourse." Hence there is a strategic reason to ask how capabilities relate to rights. But, more importantly, there is also a theoretical reason, as several capability theories express their claims in terms of people's individual entitlements. Nussbaum has most explicitly described her own version of the capabilities approach in terms of universal entitlements, and has also argued that she sees her capabilities theory as a

species of the human rights theory. Yet clearly rights are a different moral category than capabilities: how are we then to understand their relation?

Capability scholars have tended to be critical of the notion of rights, yet have also seen the moral, political, and rhetorical power of rights. The challenge is therefore to give rights a place in a capability ethics that makes that theory stronger. Why have capability scholars been critical of the notion of rights? One important risk in the rights discourse lies in the danger to conflate moral rights with legal rights. Capability ethics would have no problem with granting people moral rights to certain capabilities (possibly up to a certain level), as part of a more detailed specified capabilitarian theory. Yet the rights discourse runs the risk of overemphasizing the legal aspects of rights. The core ethical commitment of the capability approach is to secure certain freedoms and access to a certain level of quality of life to all (to the extent that this is feasible). Capabilities are thereby the ends, and rights are one of the possible means. But rights may not be enough to secure those ends, and other means may be needed in order to give people genuine access to capabilities. For example, if certain groups of people suffer from stigma or a societal taboo – as is the case with some disabilities – then rights alone will not be enough for people with those disabilities to be able to fully participate in social life and hence have access to some centrally important capabilities. Social strategies to reduce or even eliminate stigma may therefore be a much needed complement to the granting of rights in order for people to be fully able to flourish. This example also raises the question of who holds the corresponding duties to secure rights, and who holds the duties to ensure those other additional measures, such as breaking down the stigma. There is a risk with a legal rights discourse that it may induce policymakers to being contented when they have strictly followed the rules that a limited interpretation of the rights imposes on them, even when additional efforts are necessary to meeting the goal that underlies the right. And the additional capabilities-enhancing measures that are needed to complement rights may not always be seen as within the legitimate scope of government intervention, and hence we need nonstate actors to deliver those, such as civil society groups or religious organizations.

How should we characterize the role of rights within capability ethics? Rights clearly are important in daily discourse. However, at the theoretical level, rights are always rights to something. Clearly, capability ethics would want rights to target capabilities. This is also how some have believed that the capability approach can best be understood. For example, Harry Brighouse (2004: 80) writes, "It is more illuminating to think of capabilities as the bases of rights claims. If someone claims that there is a fundamental right to X, it is incumbent on them to justify it; and justification will proceed by showing how the right to X is required to serve some capability. If there is no capability that it serves, then it is not a fundamental right."

But understanding rights as capability-rights still leaves several important options open. One important question is whether such a rights-supported capabilities ethic will be consequentialist or deontological. If rights are seen as mere

instruments to serve the expansion of capabilities, then we will end up with a capability ethics that is consequentialist in nature. Such a theory could easily supplement rights with other instruments aimed at expanding capabilities (e.g., calling upon people's voluntary contributions), since the rights have a purely instrumental role.

However, if (some) capability rights are regarded as side constraints that cannot be violated no matter what, then we are entering the terrain of deontological theories. This seems to be how Nussbaum (1997: 300) understands her capabilities theory: " A list of human rights typically functions as a system of side constraints in international deliberation and in internal policy debates. That is, we typically say to and of governments, let them pursue the social good as they conceive it, so long as they do not violate the items on the list. I think this is a very good way of thinking about the way a list of basic human rights should function in a pluralistic society, and . . . I regard my list of basic capabilities this way, as a list of very urgent items that should be secured to people no matter what else we pursue." Yet despite Nussbaum's repeated reference of her list as a species of the human rights approach, there remain many unresolved conceptual questions to answer regarding capability rights.

Bernard Williams (1987: 100) was one of the first moral philosophers to point out the need for a careful conceptual and theoretical analysis of the relationship between rights and capabilities. Clearly some progress has been made (e.g., Vizard 2006; van Hees forthcoming) but this is another area where much more work is needed if we want to develop a mature capabilitarian ethics.

An Alternative to Utilitarianism?

The capability approach explicitly aims at providing an alternative to normative views that determine right and wrong simply by judging people's respective mental states (happiness, pain, etc.). This theme was present in Sen's launching of the capability approach in his 1979 Tanner Lectures (Sen 1980), and can be seen as an important move in the development of the capability approach (Qizilbash 2008: 54). Sen (1999: 59) characterizes welfarist theories as those consequentialist theories that restrict "the judgments of states of affairs to the utilities in the respective states (paying no direct attention to such things as the fulfillment or violation of rights, duties, and so on)." He rejects such theories because, whatever their further specifications, they rely exclusively on utility and thus exclude nonutility information from our moral judgments (Sen 1999: 62).

Sen is concerned not only with the information that is included in a normative evaluation, but also with the information that is excluded. The nonutility information that is excluded by utilitarianism includes a person's additional physical needs, due to being physically disabled for example, but also social or moral principles, such as human rights or the specific principle that men and women should be paid

the same wage for the same work. For a utilitarian, these features of life and these principles have no intrinsic value. Men and women, for example, should not be paid the same wage as long as women are satisfied with lower wages or total utility is maximized. But Sen believes it mistaken to think that such egalitarian and other moral principles should not be taken directly into account in our moral judgments. However, note that it is a matter of philosophical dispute whether a moral defense of basic liberties can consistently and convincingly be derived from a capabilities theory; Henry Richardson (2007) has argued that the idea of capabilities cannot well capture the social, institutional, and deontic aspects of basic liberties. If Richardson is right, then the capability approach may, perhaps, have a valid critique on the blind spots of utilitarianism, but not the answer of how to rectify these.

Thus the normative theories that Sen attacks include those that rely exclusively on mental states. This does not mean that Sen thinks that mental states, such as happiness, are unimportant and have no role to play, for they too are functionings that we sometimes have reason to value. Rather, it is the exclusive reliance on mental states that he rejects.

One could question whether the attack of Sen and some other capability scholars on utilitarianism is as successful as it may seem to them. One worry is that capability scholars attack the most simplified version of utilitarianism, or that they exaggerate the difference between (some versions of) utilitarianism and the capability approach. Based on a reading of J.S. Mill's work, Qizilbash (2008: 58) concludes that "the strong contrast which Sen sometimes makes between classical utilitarianism and his capability view is overdone." Are all versions of utilitarianism vulnerable to the capability critiques?

Note that a capability ethics which endorses a list of capabilities as fundamental human rights which function as side constraints, will use the typical role that rights as side constraints play in making an antiutilitarian point (Nussbaum 1997: 300). Given that capabilitarian theories can, but need not, endorse capability rights as side constraints, the conclusion must be that some capabilitarian theories will be further removed from utilitarianism than others.

Towards an Ethical Theory

A claim that has recurred repeatedly throughout this essay is that the capability approach is clearly a normative framework, but at its current stage of development has not yet been developed into a proper ethical theory. If one were to undertake that project, what would a capabilitarian ethical theory look like? In this final section I am to give a brief sketch of what such a theory would look like. What follows will not answer all the questions we will have to face when developing a full-blown theory, but it will give us at least a better sense of the rough nature and shape of a capability ethical theory.

Most standard accounts of ethical theory, at least in so far as we are concerned with our actions and choices rather than the formation of character, generally stipulate that ethical theory consists of two parts:

(1) The theory of value or the theory of the good, which specifies which states of affairs are intrinsically good and which are intrinsically bad.
(2) The theory of the right that specifies which actions are right and which are wrong.

Ethical theories also (tend to) specify the relationship between "the good" and "the right."

Can we apply this basic terminology to characterize the capabilities approach? Yes, we can. My proposal is to understand the capability approach as an incomplete ethical theory, consisting of the following three major elements.

The first major characteristic of the capabilitarian ethical theory is that it offers *an incomplete and underspecified theory of value*. However, while it is incomplete and underspecified, as was argued throughout this essay, it does entail the following four more precise and specific claims (1A–D).

Claim 1A: *Functionings and capabilities form the "evaluative space."* Functionings, or capabilities, or a combination of both, have ultimate value. They are the most important aspect of our account of value, and the account of value is itself very weighty in our overall ethical judgment. This proposition points to the first underspecification of the capabilities approach: when using the approach to develop a theory that has teeth, we need to decide whether we think only capabilities matter, or only functionings, or a combination of both – and for the latter, which combination of functionings and capabilities.

Claim 1B: *Not all functionings are positively valuable.* Some functionings have a negative value (Nussbaum 2003), for example, the functioning of being affected by a painful, debilitating, and ultimately incurable illness, or persistently being lonely.

Claim 1C: *Functionings and capabilities are morally neutral categories, and we need to distinguish between normatively relevant and normatively irrelevant capabilities* (Nussbaum 2003, 2011a; Robeyns 2003, 2011). The capability approach, at least in the version that Sen defends, entails a normative claim, namely, that we should focus on capabilities that people have reason to value; yet that not all capabilities are valuable. There are many beings and doings that have negative value (e.g., being raped), that have neither positive nor negative value (the capability to be able to choose between types of virtually identical washing powder), or that may be valuable for some ethical purposes, but not for others (e.g., in discussions on what the global rich owe to the global poor, we would not argue that we should be concerned with undernourished children's capability to go to the cinema; yet when we think about *relative* poverty in a western-European country, one could plausibly argue that never being able to go to the cinema can be taken to be a deprivation for such a child). This implies that we have a second

significant underspecification in the capability approach: we need to specify which capabilities matter. This specification may differ for different areas in life (e.g., politically relevant capabilities versus capabilities that are more comprehensively valuable), and also for different types of capabilitarian theories or analysis that are developed (a theory of welfare economics, an assessment of quality of life in one country or comparison of averages between countries, a theory of social justice, and so forth).

Claim 1D: *Functionings and/or capabilities are not necessarily the only elements of ultimate value. Other things may matter too.* Capabilitarian theories could endorse functionings and/or capabilities as their account of ultimate value, but may add other elements of ultimate value, such a procedural fairness. This implies that the capability approach is in itself incomplete as an account of value, since it may have to be supplemented with other elements of value. Sen (1993; 2002: ch. 20; 2009a: 27–8) has been a strong defender of (1D), for example when he argued that capabilities capture the opportunity aspect of freedom, but not the process aspect of freedom, which is also important and of value.

The second major characteristic of the capabilitarian theory of ethics is that it contains a weak proposition about the right: whenever one's actions involve a notion of the good, one should use the theory of value spelled out in what has been said so far.

The third major characteristic of the capabilitarian theory is another (weak and partial) proposition about the right: there may be views of the right that do not refer to the good that are legitimate and can complement (1) and (2). However, the capability approach does not tell us what those claims about the right are. For example, consider the following claim: "Choosing for a dictatorship as the political regime to govern a country is always wrong." This claim is a claim about the right, which does not make reference to an account of value. A more specified capabilitarian theory could endorse this claim, without making reference to people's functionings and capabilities to justify that claim. The capabilities approach as an ethical theory is thus agnostic about other aspects of the theory of the right.

This is a first attempt at spelling out how a capabilitarian ethical theory would look, and the way I have formulated it may contain confusions and need to be improved. However, my goals here have not been to develop and present a fully fleshed out and matured theory of capability ethics, but rather to illustrate that it is possible to give a general description of the capability approach as an incomplete ethical theory, which can be specified and further developed into a range of theories that rely on (parts of) an ethical theory (and may or may not also entail nonethical accounts, such as explanatory accounts). Such a project should be valuable not only for practical philosophers but also for other theorists who borrow from ethics too. For example, the eminent British welfare economist Tony Atkinson (2008) has argued that economics should be reconsidered to be a moral science. If one endorses the view that the final values that welfare economics should

target are functionings and capabilities, then one can use the above sketched ethical theory to develop a capabilitarian welfare economics.

References

Alkire, S. (2002) *Valuing Freedoms: Sen's Capability Approach and Poverty Reduction,* New York: Oxford University Press.

Anderson, E. (1999) "What is the Point of Equality?" *Ethics* 109 (2): 287–337.

Atkinson, A.B. (2008) "Economics as a Moral Science," Inaugural Joseph Rowntree Foundation Lecture, University of York.

Barclay, L. (2003) "What Kind of Liberal is Martha Nussbaum?" *SATS: Nordic Journal of Philosophy* 4 (3): 5–24.

Basu, K. (1987) "Achievements, Capabilities, and the Concept of Well-being," *Social Choice and Welfare*, 4: 69–76.

Brighouse, H. (2004) *Justice*, Cambridge: Polity Press.

Brighouse, H. and Robeyns, I., eds. (2010) *Measuring Justice: Primary Goods and Capabilities*, Cambridge: Cambridge University Press.

Carter, Ian, Kramer, Matthew, and Steiner, Hillel, eds. (2007) *Freedom: A Philosophical Anthology*, Oxford: Blackwell.

Chakraborty, A. (1996) "On the Possibility of a Weighting System for Functionings," *Indian Economic Review* 31: 241–50.

Crocker, D.A. (2008) *Ethics of Global Development: Agency, Capability and Deliberative Democracy*, Cambridge: Cambridge University Press.

Dworkin, R. (2002) *Sovereign Virtue*, Cambridge, MA: Harvard University Press.

Fleurbaey, M. (2002) "Development, Capabilities and Freedom," *Studies in Comparative International Development* 37: 71–7.

Hees, M. van (forthcoming) "Rights, Goals and Capabilities," *Philosophy, Politics & Economics.*

Kaufman, A. (2006) "Capabilities and Freedom," *Journal of Political Philosophy* 14 (3): 289–300.

Nussbaum, M. (1997) "Capabilities and Human Rights," *Fordham Law Review* 66: 273–300.

Nussbaum, M. (2000) *Women and Human Development: The Capabilities Approach*, Cambridge: Cambridge University Press.

Nussbaum, M. (2003) "Capabilities as Fundamental Entitlements: Sen and Social Justice," *Feminist Economics* 9 (2–3): 33–59.

Nussbaum, M. (2006) *Frontiers of Justice: Disability, Nationality, Species Membership*, Cambridge, MA: Harvard University Press.

Nussbaum, M. (2011a) *Creating Capabilities: The Human Development Approach*, Cambridge, MA: Harvard University Press.

Nussbaum, M. (2011b) "Political Liberalism," *Philosophy and Public Affairs* 39 (1): 3–45.

Pierik, R. and Robeyns, I. (2007) "Resources versus Capabilities: Social Endowments in Egalitarian Theory," *Political Studies* 55 (1): 133–52.

Pogge, T. (2002) "Can the Capability Approach Be Justified?" *Philosophical Topics* 30 (2): 167–228.

Qizilbash, M. (2008) "Amartya Sen's Capability View: Insightful Sketch or Distorted Picture?" in *The Capability Approach. Concepts, Measures and Applications*, eds. F. Comim, M. Qizilbash, and S. Alkire, Cambridge: Cambridge University Press, pp. 53–81.

Richardson, H.S. (2007) "The Social Background of Capabilities for Freedoms," *Journal of Human Development* 8 (3): 389–414.

Robeyns, I. (2003) "Sen's Capability Approach and Gender Inequality: Selecting Relevant Capabilities," *Feminist Economics* 9 (2–3): 61–92.

Robeyns, I. (2011) "The Capability Approach," in *Stanford Encyclopedia of Philosophy*, online at http://plato.stanford.edu/entries/capability-approach/ (accessed February 3, 2012).

Schokkaert, E. (2007) "Capabilities and Satisfaction with Life," *Journal of Human Development* 8 (3): 415–30.

Sen, A. (1980) "Equality of What?" in *Tanner Lectures on Human Values*, vol. 1, ed. S. McMurrin, Cambridge: Cambridge University Press, pp. 196–220.

Sen, A. (1985a) *Commodities and Capabilities*, Amsterdam: North-Holland.

Sen, A. (1985b) "Well-Being, Agency and Freedom: The Dewey Lectures 1984," *Journal of Philosophy* 82 (4): 169–221.

Sen, A. (1987) "The Standard of Living," in *The Standard of Living*, ed. G. Hawthorn, Cambridge: Cambridge University Press, pp. 1–38.

Sen, A. (1992) *Inequality Re-examined*, Oxford: Clarendon Press.

Sen, A. (1993) "Capability and Well-Being," in *The Quality of Life*, eds. M. Nussbaum and A. Sen, Oxford: Clarendon Press, pp. 30–53.

Sen, A. (1999) *Development as Freedom*, New York: Knopf.

Sen, A. (2002) *Rationality and Freedom*, Cambridge, MA: Harvard University Press.

Sen, A. (2005) "Human Rights and Capabilities," *Journal of Human Development* 6 (2): 151–66.

Sen, A. (2009a) "Capability: Reach and Limit," in *Debating Global Society. Reach and Limit of the Capability Approach*, ed. Enrica Chiappero-Martinetti, Milan: Feltrinelli, pp. 15–28.

Sen, A. (2009b) *The Idea of Justice*, London: Allen Lane.

Vallentyne, P. (2005) "Debate: Capabilities versus Opportunities for Wellbeing," *Journal of Political Philosophy* 13: 359–71.

Vizzard, P. (2006) *Poverty and Human Rights: Sen's 'Capability Perspective' Explored*, Oxford: Oxford University Press.

Williams, A. (2002) "Dworkin on Capability," *Ethics* 113: 23–39.

Williams, B. (1987) "The Standard of Living: Interests and Capabilities," in *The Standard of Living*, ed. G. Hawthorn, Cambridge: Cambridge University Press, pp. 94–102.

Wolff, J. and de-Shalit, A. (2007) *Disadvantage*, Oxford: Oxford University Press.

Chapter 20

Feminist Ethics

Alison M. Jaggar

Throughout the history of Western ethics, the moral status of women has been a persistent though rarely central topic of debate. A few isolated voices have contended that women are men's moral equals but most of the dominant figures in the tradition have offered ingenious arguments to justify women's subordination to men. Despite the long history of this controversy, the expression "feminist ethics" was coined only in the 1980s, after feminism's "second wave" had swept into the academies of North America and, to a lesser extent, western Europe – a critical mass of philosophers for whom the status of women was an important ethical concern. The appearance of this expression not only signaled a perception that attention to women and gender was indispensable to adequately understanding many issues in practical ethics; it also reflected a new belief that women's subordination had far-reaching, though hitherto unnoticed, consequences for ethical theory.

Feminist ethical theory is distinguished by its exploration of the ways in which cultural devaluation of women and the feminine may be reflected and rationalized in the central concepts and methods of moral philosophy. Not all feminist philosophers are convinced that Western ethical theory is deeply flawed by such devaluation; on the contrary, some propose that one or another existing theory – perhaps with a little fine tuning – is entirely adequate to address feminist ethical concerns. However, many feminist philosophers contend that Western ethical theory is deeply male biased. Although they sometimes disagree with each other regarding the nature of this alleged bias and/or in their prescriptions of an alternative to it, their work is characterized by attention to certain recurrent themes. The present essay traces the evolution of those themes and in so doing offers a critical reconstruction of the development of Western feminist ethical theory.

The Blackwell Guide to Ethical Theory, Second Edition. Edited by Hugh LaFollette and Ingmar Persson.

Including Women in Ethical Theory

Most of the great Western philosophers assigned a higher ethical priority to men's interests than to women's, contending that women's proper role was to support men in men's undertakings. One theme continuing from ancient to modern times is that women's primary responsibility is to produce children for their husbands and the state, while providing their husbands with physical and emotional care. Aristotle, for example, asserts that a wife must obey and serve her husband because he has bought her with a great price; Aquinas writes that woman was made to be a helper to man, "not indeed, as a helpmate in other works, as some say, since man can be more efficiently helped by another man in other works; but as a helper in the work of generation," and Rousseau argues that "woman is intended to please man." Feminist philosophers have revealed what Susan Okin calls "functionalist" treatments of women by, among others, Plato, Aristotle, Aquinas, Hobbes, Locke, Rousseau, Kant, Hegel, Nietzsche, and Rawls (Okin 1979, 1989; Clark and Lange 1979).

Even though Western philosophers generally treated women's interests as instrumental to men's, they regarded this treatment as standing in need of justification; their justifications typically took the form of arguing that women were in some important sense less fully or perfectly human than men. Some held that women were incapable of the same moral perfection as men: for instance, Aristotle says that women's temperance, courage, and justice are of a different – and lesser – kind than men's; Rousseau asserts that women's merit consists in such "feminine" virtues as obedience, silence and faithfulness; Kant writes, "The virtue of a woman is a *beautiful virtue*. That of the male sex should be a *noble virtue*." Many philosophers argued that women's capacity for reason was also different from and inferior to men's; major figures developing such arguments included Aristotle, Aquinas, Rousseau, Kant, Hegel, Nietzsche, and Sartre. Since the Western tradition typically regards rationality as the essential human characteristic, often defining moral agency in terms of the capacity for reason, arguments that women's reason is inferior to men's are deeply damaging to women's aspirations for equality. They suggest that women may be less morally valuable than men because their supposedly lesser rationality places them closer to animals and further from God; moreover, by entailing that women have less moral authority than men, they provide a strong rationale for placing women under men's political authority.

Including Women Equally as Objects of Ethical Concern

At the start of the twenty-first century, when a commitment to women's equality is enshrined in United Nations declarations of human rights as well as in many

national constitutions, it may seem hardly controversial to claim that women's interests should weigh equally with men's. Yet despite the lip service paid almost universally to the idea that persons should receive equal moral consideration regardless of their sex, feminists note that in practice public policies often accord less weight to women's interests than to men's. Sometimes this inequality of consideration may be attributed to faulty applications of ethical theory but sometimes feminists trace it to bias endemic in the theory itself.

Utilizing "Gender" in Ethical Analyses

One reason for public policy's frequent bias against women is that equality of consideration is often assumed to require treating men and women indistinguishably. Deliberately ignoring distinctions of sex often has the consequence that ethical analyses fail to take account of morally salient differences between men and women.

Feminist research has revealed that many superficially sex-neutral issues in fact affect men and women differently and feminists insist that these differences must be addressed by any public policy that is ethically adequate. Examples of such differences abound; for instance, women often suffer more than men from war, even though men constitute most of the combatants. Over the last century, as the proportion of civilian casualties has multiplied, women's share of the suffering has increased, since women who are not injured or killed directly are often displaced and become refugees; even in times of so-called peace, women suffer disproportionately from the allocation of tax money to military expenditures rather than to social services and benefit least from job opportunities in the military and related industries. Many issues of global justice have significantly different implications for men and women. They include: population policies that target women's rather than men's fertility; economic development policies that invest in men's enterprises while failing to acknowledge the value of women's agricultural and domestic work; foreign investments in industries that exploit women's labor; and the increasing economic prominence of the global tourism industry and its concomitant sex trade.

The above examples illustrate that men and women are differently situated in all known societies; they are subjected to systematically different norms and expectations that govern virtually every aspect of their lives. All known societies assign different work to biological males and females, different family responsibilities, different standards of appropriate sexual behavior, dress, and diet, even different norms of physical deportment and patterns of speech. To distinguish these sets of social norms and expectations from biological differences between men and women, Western feminists of the late 1960s appropriated the hitherto grammatical term "gender." They contended that, whereas sex differences were socially invariant, gender differences varied both among and within societies; they observed that masculinities and femininities, the social meanings assigned to being male and female, differed both in different societies and also among individuals of different

castes, classes and ethnicities in any given society. More recent work in feminist theory has challenged the apparent clarity of the sex/gender distinction, especially the supposed naturalness and immutability of sex, but I shall not pursue that discussion here.

The realization that gender is a variable salient for much ethical practice has convinced some feminists that it is also a category indispensable to ethical theory. Those who hold this view argue that ethical theory cannot remain satisfied with conceptualizing humans on such a high level of abstraction that their inevitable differences, including their gender differences, become invisible. They contend that an adequate ethical theory cannot conceptualize human beings as undifferentiated, ignoring gender and related characteristics such as age, ability, class, and race; instead, it requires a more complex conceptual apparatus that reflects the inevitable differences among people.

In opposition to this contention, other feminists object that what is required for more adequate analyses in practical ethics is not that ethical theory be revised but simply that those utilizing the theory take more account of the morally salient differences among individuals. Liberal feminists, in particular, often fear that elevating gender to the status of a concept in ethical theory would abandon feminism's traditional insistence that there exist no morally significant differences between men and women and so play dangerously into antifeminist hands. These liberals endorse the older feminist position that sex differences should be conceived simply as "accidental" or inessential properties qualifying an underlying – and sex-neutral – human essence; they contend that ethics should address issues of gender on the level of first-order practice rather than second-order theory. Later in this essay, we shall see how this dispute has developed.

Expanding the Domain of Ethics

Modern, although not ancient, moral philosophy has given little attention to many issues of special concern to women, most notably issues of sexuality and domestic life. This neglect has often been rationalized by a theoretical bifurcation of social life into a public domain, regulated by universal principles of right, and a private domain, in which varying goods may be properly pursued. Even philosophers like Aristotle, Hegel, and Marx, who regard the home as having some ethical importance, portray it as an arena in which the most fully human excellences are incapable of being realized.

Inspired by the 1960s slogan, "The personal is political" (and, by extension, ethical), many feminists have challenged not only philosophers' neglect of the gendered aspects of most ethical issues but also their theoretical rationale for excluding some issues altogether. Feminists point out that the public/private dichotomy is covertly gendered, since women traditionally have been excluded from what is conceptualized as the public and restricted to what is defined as the private; the home, for instance, has become symbolically associated with the feminine, despite the fact that heads of households are paradigmatically male. Feminists argue that excluding the domestic realm from the moral domain is not only

arbitrary but also covertly promotes masculine interests; for instance, by denying the conceptual resources for raising questions about the justice of the domestic division of labor, it obscures the social necessity and arduousness of women's work in the home; moreover, by relegating intimate relationships to the domain of the personal or subjective, it screens and may even license the domestic abuse of women and girls.

Contemporary feminists have sought to expand the domain of ethics to embrace not only the domestic sphere but also many other aspects of social life. They have raised ethical questions concerning abortion; sexuality, including compulsory heterosexuality, sexual harassment and rape; representation, including mass media and pornographic portrayals of women; self-presentation, including body image and fashion; and the role of language in reinforcing as well as reflecting women's subordination. Although mainstream ethics has given little attention to these issues until very recently, they all have ethically significant consequences for women's lives and are sometimes matters of life and death.

Although they may sometimes speak of including "women's issues" within the domain of ethics, feminists' use of this language does not imply that they recognize a category of women's ethical issues that is distinct from men's, much less from human, issues. What are often categorized as women's issues are also in practice men's, since men's and women's lives are always enmeshed with each other; for instance, whether or not childcare or abortion is available significantly affects the lives of men as well as women. Men are involved in domestic, sexual, and personal relations, just as women are involved in the economy, science, and the military, despite the symbolic casting of the former as feminine and the latter as masculine. Most contemporary feminists contend that, if women are more preoccupied with or affected by certain matters than men, this is not natural or inevitable but instead reflects women's culturally assigned confinement to and/or responsibility for some areas of life and their relative exclusion from others.

In order to give due weight to women's interests, many feminists assert that ethical theory must operate with a more complex set of categories and virtually all agree that it must expand its domain.

Including Women Equally as Moral Subjects

The question of moral rationality and subjectivity is logically independent of the question of moral considerability; there is no logical reason why the interests of children, mentally disabled persons, animals, or ecosystems should not count as morally equal to those of rational moral agents. However, Western disregard for women's interests has often been justified by denying that women are full moral agents and so it has often been thought necessary to validate women's moral subjectivity in order to demonstrate that women's claims to moral concern are equal with men's. Demonstrating that women should have equal political rights certainly requires establishing that they have equal moral authority.

Efforts to establish that women are full moral subjects long predate the emergence of contemporary feminist ethics. In the *Republic* (written in the fifth century BC), Plato declares that some women are capable of being guardians or rulers; in *The Book of the City of Ladies* (1405), Christine de Pisan argues that women are equal or even superior to men in such virtues as wisdom, courage, prudence, constance, and chastity; in *A Vindication of the Rights of Woman* (1792), Mary Wollstonecraft denies the existence of virtues specific to one sex or the other and insists that women are as potentially rational and as fully human as men; in *The Subjection of Women* (1869), John Stuart Mill suggests that women's apparent inferiority in reasoning and principled morality is most likely due to their different socialization; and early in the twentieth century Bertrand Russell argued that women's intelligence and virtue varied in just the same ways as men's.

At the beginning of the twenty-first century, contending that women are moral subjects equally with men may seem as superfluous as arguing that women are entitled to the same moral consideration. But although women now vote in all Western democracies, their suffrage was achieved in many of these nations only during the lifetimes of many people still living today. British women received the vote after World War I but they did not receive it on the same terms as men until after World War II; French and Italian women received the vote only after World War II; Swiss women were unable to vote in national elections until 1973 and did not have suffrage in all cantons until the 1990s. The dearth of women political leaders suggests that Western publics still lack confidence in women's moral authority. Although women's potential for moral subjectivity is rarely disputed directly nowadays by respected authorities, recent moral psychology has claimed that women are less likely than men to actualize that potential and attain the highest levels of moral development (Kohlberg 1981). During the 1980s, however, some theorists altered feminism's traditional response to such claims: instead of continuing to insist that women were capable of reaching men's level of moral development, they began to challenge the standard by which moral rationality and subjectivity were judged. The following sections explore their challenges.

Is Modern Ethical Theory Male Biased?

When Western feminists criticize modern ethical theory, their usual targets are those liberal theories, rooted in the European Enlightenment, that still dominate contemporary Western philosophy. Such theories include Kantianism and its descendants such as some versions of contractarianism and discourse ethics, utilitarianism in its various forms and, sometimes, existentialism. Few feminists wholeheartedly endorse neo-Aristotelian theories such as communitarianism and virtue ethics but their reservations about these theories so far have received less

development than their reservations about liberal theory. This may be partly because their criticisms of modern liberal theory share common elements with neo-Aristotelian criticisms.

By setting aside traditional constraints on the realm of the ethical and paying attention to gender differences, some feminists have succeeded in utilizing liberal theory to illuminate a number of practical ethical issues of special concern to women; for example, Susan Okin uses Rawlsian contractarian theory to show how contemporary marriage practices discriminate against women (Okin 1989). Despite such achievements, many feminists argue that modern ethical theory is so thoroughly infected with masculine bias that it has only limited usefulness for feminism. The present section of this essay elaborates feminist criticisms of modern ethical theory and the next section, "Women's Experiences as a Paradigm for Ethical Theory," outlines an influential feminist alternative; in the following section, "Ethical Theory: Feminine or Feminist?" I offer some critical discussion of that alternative.

The Values of Modern Theory as Ethically Inadequate

Despite their differences, Enlightenment ethical theories have much in common; most fundamentally they share a commitment to the equal moral value of every human individual. In the Kantian tradition, this value is expressed by recognizing the worth of each individual's autonomy; in the utilitarian tradition, it is expressed by assigning equal weight to each individual's happiness. In both traditions, realizing this value requires nonpaternalism expressed by noninterference in the lives of others (Baier 1987).

Few feminists reject modern ethical values entirely and some have deployed them to good effect, arguing that women are entitled equally with men to respect and autonomy. However, even when feminists endorse modern values, they often propose that widely accepted interpretations of them should be revised; for instance, some fault common interpretations of Kantian theory for assuming that autonomy is a natural property possessed by all normal adults instead of recognizing that it is a potential realizable only in community.

A more fundamental challenge to modern ethical theory is the charge that it often generates ethical prescriptions that, according to its critics, are morally repellent to many women. These critics do not attribute the alleged incompatibility between ethical theory and women's moral sensibilities to improper application of the theories, still less to deficient sensibilities in women. On the contrary, they assert that liberal values offer an impoverished ethical vision, providing a model of human interaction that at best is appropriate only for a limited domain of life and at worst may rationalize inhumanity to others. For instance, Baier notes that "noninterference can, especially for the relatively powerless, such as the very young, amount to neglect, and even between equals can be isolating and alienating" (Baier 1987: 48–9).

Impartiality is a core value in modern ethical theory but, since about 1980, it has been challenged both by communitarians and some feminists. Their criticisms overlap but are not identical. The ideal of impartiality requires that each individual receive equal consideration, regardless of an agent's subjective connections or loyalties to particular individuals. Some feminists have argued that this ideal is unrealizable, since it is psychologically impossible for human thinking to be detached from its context of origin or from its motivating passions and commitments (Noddings 1984; Young 1990: 103–5). Others, communitarians as well as feminists, have argued that the ideal of impartiality is morally defective, since it entails readiness to sacrifice those we love to abstract principles and absent strangers. Some argue that treating people as ethically equivalent denies the moral significance of individuality, which appreciates precisely the uniqueness of each person (Sherwin 1987). According to its feminist critics, too much emphasis on impartiality underrates those personal values that are more fundamental to a good human life.

Modern Conceptions of the Moral Subject as Unrealistic and Repellent

Modern ethical theory typically utilizes a neo-Cartesian conception of the moral subject as an agent that is essentially rational. Although the canonical theorists certainly assumed that moral agents were embodied members of communities, they regarded people's bodies and community memberships as "accidental" or contingent properties irrelevant to their claims to moral subjectivity.

Some feminists have found that the modern conception of the subject is a valuable resource for maintaining that women are full moral agents, disqualified neither by their female bodies nor by their frequently dependent social status. Accepting the modern conception, they insist that women are just as capable as men of transcending the limitations of their bodies and they argue that the Western philosophical association of men with mind and women with body has no defensible basis. In their view, this association serves simply to rationalize men's political dominance, as well as social arrangements that assign to women the primary responsibility for taking care of bodily needs.

Other feminists are critical of modern ethical theory's abstract, rationalistic, and individualistic conception of the moral subject. These critics often focus on the modern devaluation of the body, charging that it has been an important contributor to what they perceive as the flaws in Enlightenment ethical theory. They argue that devaluing the body in comparison with the mind has encouraged ethical theory to ignore many fundamental aspects of human life and to posit ideals unattainable by human beings. Disparagement of the body, they contend, turns theoretical attention away from bodily related differences among individuals, such as age, sex, and ability, and encourages regarding people as indistinguishable and interchangeable. Ethical reflection on embodiment would reveal that inequality, dependence and interdependence, specificity, social embeddedness, and historical

community must be recognized as permanent features of human social life and that seeking to transcend these is a waste of time. Instead of devoting so much ethical attention to abstractions such as equality, autonomy, generality, isolated individuals, ideal communities, and the universal human condition, many feminists argue that ethical theory should pay more attention to people's bodies. This would enable it to recognize the central ethical issues of vulnerability, development, and mortality rather than changelessness, of temporality and situatedness rather than timelessness and nonlocatedness, of particularity rather than universality, and of interdependence and cooperation rather than independence and self-sufficiency.

Modern Conceptions of Moral Rationality as Unrealizable or Pathological

Enlightenment ethical theory regards rationality both as a natural property belonging to all normal human adults and as the only reliable guide to distinguishing right from wrong action. Viewing emotions as contaminants of pure reason, it defines moral rationality in terms of individuals' ability to consider dispassionately the interests of all those affected in any situation, thus overcoming the supposedly normal human tendency towards self-interested bias. Some feminists dispute both the descriptive and the prescriptive elements of this account. On the descriptive level, they challenge the assumption that people are predominantly self-aggrandizing, an assumption they see as facilitated by liberalism's disregard for human embodiment. Instead, they contend that the social meanings attached to bodily characteristics such as parentage, age, or sex, result in embodied individuals developing moral identities that are not purely abstract and universal but also defined by the social relations involved in the meanings assigned to various specific bodies. Individuals with relational moral identities are unlikely to make a sharp separation between their own interests and those of others; they are more likely to be moved by considerations of particular attachment than by abstract concern for duty, more by care than by respect, and more by responsibility than by right.

For feminist critics of modern ethical theory, people's propensities to care for others and to regard their own interests as linked with those of others are not just weaknesses to be overcome by moral reason. Baier challenges what she calls the rationalism or intellectualism of modern moral theory, a rationalism that assumes that we need not worry what passions persons have, as long as their rational wills can control them.

> This Kantian picture of a controlling reason dictating to possibly unruly passions also tends to seem less useful when we are led to consider what sort of person we need to fill the role of parent, or indeed want in any close relationship. It might be important for father figures to have rational control over their violent urges to beat to death the children whose screams enrage them, but more than control of such nasty

> passions seems needed in the mother or primary parent, or parent substitute, by most psychological theories. They need to love their children, not just to control their irritation.
>
> (Baier 1987: 55)

We shall see in the next section, "Women's Experience as a Paradigm for Ethical Theory," that not only do some feminists deny that emotions are necessarily subversive of moral reason; they regard them as indispensable to it. In modern ethical theory, impartiality is not only a substantive ideal but also a defining characteristic of moral rationality, providing a necessary and sometimes sufficient condition of right action. We have seen already that some feminists challenge the substance of this ideal; others may accept the ethical intuition at its core but observe that the concept is too indeterminate to guide right action. Modern moral philosophers have offered a variety of recommendations for achieving impartiality, such as disregarding one's own self-interested motivations or adopting others' points of view, but a number of feminist critics have argued that these recommendations are quite unhelpful since they cannot be operationalized in practice. For instance, Marilyn Friedman notes that the limited nature of individuals' experience and of their familiarity with the thinking of others makes it highly unlikely that any real person (as opposed to an archangel) could project herself imaginatively into the standpoint of another, let alone of many others; nor could one who attempted this imaginative feat ever know how far she had been successful. Friedman concludes that available philosophical conceptions of impartiality offer no practical guide to moral justification. She recommends that people who wish to do the right thing should focus instead on partiality, concentrating on eliminating particular nameable biases from their thinking (Friedman 1993: 31).

The Alleged Masculinity of Modern Ethical Theory

Why do some feminists allege that the distinctive values of modern ethical theory, its conception of the moral subject, and its conception of moral reason are characteristically masculine? What is specifically masculine about valuing equality, autonomy and respect, understanding human subjects in terms of their minds rather than their bodies, and construing moral reason in terms of dispassionate impartiality? Marxist critics have long argued that modern ethical theory is based on a "possessive individualist" conception of human nature that portrays humans as essentially separate from others, insatiably appetitive, and with interests typically in conflict, and they have charged that this conception reflects the adversarial market relations of bourgeois society. Feminists have accepted much of this picture but they have added the claim that men are more likely than women to understand human nature in such adversarial terms (Gilligan 1982). Few feminists attribute this alleged difference in perspective to some innate psychological differences between the sexes; instead, they explain it by reference to the contingently

different social situations of men and women. Some draw on neo-Freudian object relations theory, which appeals to gendered patterns of parenting to argue that a preoccupation with separation is distinctively masculine. Others argue that disregard of the body is a luxury available only to those whose bodies are normative and/or who are freed from primary responsibility for bodily maintenance.

Basing ethical theory on a model of human nature that reflects men's distinctive experiences and values is problematic most obviously because it valorizes the ethical perspectives of only one segment of the population. Feminists further contend that the dominant model fails to describe accurately the moral psychology not only of most women but also of many men. Basing ethical theory based on false empirical postulates is likely to result in unrealizable ideals and epistemologies. Moreover, an ethical theory based on a masculine image of human nature devalues the symbolically feminine dimensions of human life; it also neglects more "feminine" ethical visions, promoting an image of the ethical life that many find repellent, especially many women. In addition to advancing an exclusionary, limited, and – to many – repugnant ethical vision, modern ethical theory impugns the moral authority of those who disagree with it by labeling them as morally deviant, immature, or irrational (Gilligan 1982). For its feminist critics, modern ethical theory proposes a male-biased ethical vision that justifies itself by an equally male-biased moral epistemology.

Some feminists charge that modern ethical theory is masculine, finally, in projecting its devaluation of women and of feminine experience onto the universe at large. It follows the larger Western philosophical tradition that interprets reality through conceptual dichotomies such as culture/nature, transcendence/immanence, permanent/unchanging, universal/particular, mind/body, reason/emotion, and public/private. By associating the more highly valued term with masculinity and the less valued with femininity, Western ethical theory inscribes cultural hostility for women into its portrayal of ultimate reality.

Women's Experience as a Paradigm for Ethical Theory

In response to the charge that modern ethical theory assumes masculine experience as normative, some feminist ethics has sought to take women's experience as its paradigm or at least as its point of departure. The best-known example of this approach is the ethics of care, which elaborates a moral perspective said to arise from women's characteristic experiences of nurturing particular others, especially their experiences of rearing children (Gilligan 1982; Noddings 1984; Ruddick 1989; Held 1993). Although the project of deriving ethics from women's experience is generally associated with the ethics of care, a few feminists reject care's emphasis on nurturing or mothering and seek to derive ethics from other facets of women's experience; for instance, Sarah Lucia Hoagland aims to derive new value from reflecting on lesbian lives (Hoagland 1988).

Since feminist ethical theory is often identified with the ethics of care, it is worth emphasizing that neither the ethics of care nor the project of basing ethical theory exclusively or primarily on women's experience should be taken as feminist orthodoxy. I have nevertheless chosen to devote considerable space to care ethics because it offers the best-known and, many believe, most radical challenge made by feminists to modern ethical theory. It contends that attention to women's moral experience advances values that are ethically superior to those characteristic of modernity and fosters more adequate conceptions of moral subjectivity and moral rationality.

Appreciating the Values Implicit in Women's Ethical Practice

Proponents of care ethics characteristically advocate that ethical priority should be given to the values that they see as central to women's practices of nurturing and especially of mothering; these include the values of emotional sensitivity and responsiveness to the needs of particular others, intimacy and connection, responsibility and trust. Modern ethical theory has always feared that justice would be subverted if too much weight were accorded to these values, but it has accepted them in what it has seen as their proper place, namely, within the limited domain of intimate personal relations; on the epistemological level, it has accorded them a similarly minor role as possible motivators to right action. Most care theorists reject this relegation to what Benhabib calls "the margins of ethical theory"; instead, they often propose that the values hitherto associated with the private domain should become more prominent both in ethical theory and in society at large. For instance, Virginia Held considers how to export to wider society the relations suitable for mothering persons and children (Held 1993). Sara Ruddick considers how "maternal thinking" may promote a politics of peace (Ruddick 1989). Joan Tronto argues that care may be a political as well as an ethical ideal, describing "the qualities necessary for democratic citizens to live well together in a pluralistic society" (Tronto 1993: 161–2).

"Feminizing" the Ethical Subject

We have seen that modern ethical theory is dominated by a neo-Cartesian model of the subject as disembodied, asocial, unified, rational, and essentially similar to all other selves; we have also seen that some feminists accept this model but that many challenge it. In developing their challenges, feminists have drawn insight from several traditions, such as Marxism, psychoanalysis, communitarianism, and postmodernism, but they have been especially influenced by the work of psychologists such as Jean Baker Miller and Carol Gilligan. Gilligan asserted that women and girls tend to see themselves as connected to others and to fear isolation and

abandonment, unlike men who are said to see themselves as separated from others and to fear connection and intimacy. She reported that women's conception of their selves as relational gives them different moral preoccupations and encourages them to construe moral dilemmas as conflicts of responsibilities rather than rights, to seek to resolve those dilemmas in ways that will repair and strengthen relationships, to practice positive caretaking rather than respectful nonintervention, and to prioritize the personal values of care, trust, attentiveness, and love for particular others above impersonal principles of equality, respect, rights, and justice. Many feminist ethical theorists advocate a so-called relational model of the self. They contend that such a model is superior to the Cartesian conception for understanding not only women but also men; contrary to the view of human nature presupposed by modern ethical theory, all human beings in fact are interdependent, constrained, and unequal. Thus some feminists argue that a relational conception of moral subjectivity is more adequate empirically than an atomistic model and also generates a more acceptable ethics (Whitbeck 1984). For these theorists, "masculine" consciousness is false consciousness.

Rethinking Moral Rationality

The "style" of moral reasoning associated with care ethics is often contrasted with that characteristic of justice ethics. Whereas justice thinking focuses primarily on the structure of an ethical situation, deliberately disregarding the specific identities of the individuals involved, care thinking is characterized by a distinctive ethical orientation toward particular persons. This orientation has both affective and cognitive dimensions: caring individuals are both concerned about the other's welfare and perceive insightfully how it is with the other. Contrary to justice thinking, which is portrayed as appealing to universalizable moral principles that guide impartial calculation of who is entitled to what, accounts of care thinking emphasize its responsiveness to particular situations whose morally salient features are perceived with an acuteness thought to be made possible by the carer's emotional posture of empathy, openness, and receptiveness (Blum 1992).

Perhaps the most distinctive and controversial feature attributed to care thinking is its particularity; this means not only that it addresses the needs of others in their concrete specificity but that it is unmediated by general principles. Care responds to others as unique, irreplaceable individuals rather than as "generalized" others seen simply as representatives of a common humanity (Benhabib 1992). Such responsiveness requires paying as much attention to the ways in which people differ from each other as to the ways in which they are the same. Another aspect of care's particularity is that its conclusions are nonuniversalizable; that is, they carry no implication that someone else in a similar situation should act similarly. The radical particularism of care thinking challenges a fundamental assumption of modern ethical theory, namely, that appraising particular actions or practices requires appeal to general principles.

Proponents of care ethics resist reducing care to a simple emotional response; they consider it not simply as a motivator to right action, the latter determined through a process of rational calculation, but also as a distinct moral capacity with cognitive dimensions necessary to determining what actions are morally appropriate (Blum 1992). Care is not rational in the senses of being egoistic, dispassionate, or deductive, but Nel Noddings asserts that "rationality and reasoning involve more than the identification of principles and their deductive application" (Noddings 1990: 27). Proponents of care thinking regard care as rational in the broad sense of being a distinctively human way of engaging with others; it is both ethically valuable in itself and it tends to produce morally appropriate action.

Ethical Theory: Feminine or Feminist?

The ethics of care has revealed some serious gaps and biases in modern ethical theory, many of which are attributable to that theory's exclusion of women's experience and concerns. A more adequate ethical theory must, in my view, develop some means of including the moral perspectives of women, as well as the perspectives of other devalued or marginalized groups. Nevertheless, I find that the way the ethics of care so far has developed ethics from the perspective of women is problematic both in methodological principle and in ethical practice.

Can Ethical Theory Be Built on Women's Experience?

Attempts to derive ethical theory from empirical experience reflect the naturalist conviction that philosophical ideals must be compatible with people's actual moral sensibilities; on this view, apparent divergence between ethical theory and ethical practice may not be dismissed immediately as a failure in practice. Moreover, an ethical theory that is responsive to feminist concerns requires that specific attention be paid to women's ethical experience in order to acknowledge women's hitherto devalued capacities as moral agents.

Although these contentions are, in my view, correct, it is necessary to remember that naturalistic approaches to ethical theory involve characteristic moral dangers. One is conventionalism, which takes accepted values and ways of thinking as self-justifying; linked with conventionalism is relativism, which asserts that what is morally permissible varies for different moral communities. Both conventionalism and relativism are problematic for feminism, because they conflict with its steadfast opposition to all forms of male dominance.

In addition to its moral dangers, ethical naturalism faces considerable methodological problems. One of these is that the term "ethical experience" is so broad that it is unclear how it should be investigated. Another is that what people

say about ethics is notoriously unreliable as a guide to their actions. Moreover, it is difficult to find empirical confirmation for generalizations about the moral experience of large and diverse groups, such as women or lesbians, even when these generalizations are made by philosophers who themselves are women or lesbian.

Methodological problems underlie many feminist debates about how women's ethical experience should be characterized and they emerge with special clarity in the ethics of care. We have seen that care theorists assert that culturally feminine experiences such as nurturing provide the basis for an ethical vision quite distinct from that promoted by modern ethical theory. In a complex modern society, however, all unqualified generalizations about men's and women's experiences are *prima facie* dubious; the life situations of both women and men in contemporary Western societies vary so widely by class, race/ethnicity, and even generation, that it seems quite unlikely that all or most women share a moral perspective different from that of all or most men. In fact, investigations into the empirical validity of care theorists' claims have often failed to confirm a link between gender and caring; when subjects are matched for education and occupation, women often achieve almost identical scores with men on justice-oriented tests of moral development, leaving women who work in the home as the main female representatives of the care perspectives. Moreover, many men as well as women have been found to employ care thinking, especially lower-class men and men of color. For these reasons, Marilyn Friedman argues that the ethics of care is feminine in a sense that is more symbolic or normative than empirical; rather than reflecting empirical dispositions in women toward empathy, sensitivity, and altruism, she suggests that care expresses the cultural expectation that women be more empathic, sensitive, and altruistic than men (Friedman 1993: 123–4).

Recent advocates of an ethics of care acknowledge that some women think in terms of justice and some men in terms of care, but they nevertheless associate caring with women because they regard the care perspective as emerging from forms of socialization and practice that, in contemporary Western society, are predominantly feminine; these include raising children, tending to the elderly, maintaining a supportive home environment, and nursing. Joan Tronto argues that the ethics of care is associated not only with gender, but also with race and class. She links the ethical perspective of care with the work of maintaining and cleaning the body, tasks that in Western history have been relegated primarily to women but not to all women or to women exclusively; such caring work is done not only by women but also by the working classes and especially, in much of the West, by people of color (Tronto 1993). Tronto's analysis of the social genesis of care thinking fits well with Lawrence Blum's argument that justice ethics expresses a juridical-administrative perspective that is indeed masculine but which reflects the concerns not of all men but specifically of those in professional and administrative classes (Blum 1982). Together, Tronto's and Blum's arguments suggest that both the ethics of justice and the ethics of care are not only gendered but simultaneously raced and classed.

Is Women's Moral Experience a Dependable Basis for Feminist Ethical Theory?

In the preceding section, "Women's Experience as a Paradigm for Ethical Theory," we noted some difficulties in determining just what is women's moral experience. But even if we grant that the ethics of care is in some sense *feminine*, this would not be sufficient to establish it as an ethics that is *feminist*, since feminism is often critical of the feminine. One necessary condition of an ethical theory's being feminist is that it should provide conceptual resources adequate for criticizing all forms of male dominance, and some feminists, including myself, doubt that the ethics of care offers such resources.

One concern raised by a number of feminist philosophers is that the ethics of care is insufficiently suspicious of the characteristically feminine moral failing of self-sacrifice. Arguing that care for one's abuser, for instance, may be morally pathological rather than virtuous, and noting that Noddings justifies the responsibility to care for oneself only in the instrumental terms of maintaining one's capacity to care for others, some feminists have characterized care as a slave morality (Card 1990).

Other problems result from care's characteristic focus on the specific needs of particular individuals. The morally problematic situations described by care theorists typically involve only a few individuals and typically require the agent to respond to others perceived in their concrete particularity. A number of critics have wondered how this model of moral rationality can avoid partiality to the particular others known to the agent. They have also questioned how care thinking can address large-scale social or global problems involving large numbers of people who could never be known personally by any single agent.

I have worried that care thinking may distort our understanding of some morally problematic situations. Care's narrow focus is valuable in encouraging awareness of moral complexity and individual responsibility in small-scale situations but it may well obscure perception of the macro-situations that provide the context for individual encounters. For instance, it may enable us to discern insensitivity or bullying on the part of particular individuals while diverting moral attention away from the social structures of privilege that legitimate their behavior. Similarly, attending to an individual's immediate needs for food, shelter, comfort, or companionship may distract us from moral scrutiny of the structures that create those needs or leave them unfulfilled. Thus care thinking may encourage what are sometimes called band-aid or social work approaches to moral problems, rather than encouraging efforts to address them institutionally or even to prevent their occurrence through social reform (Jaggar 1995).

A final problem that I find in the ethics of care is its lack of guidance in determining which caring responses are ethically appropriate. Most care theorists acknowledge the need to distinguish appropriate from inappropriate caring but they seem to assume that this distinction is self-evident or at least that the carer/

cared-for dyad can be relied on to make it. However, such an assumption is evidently unwarranted; examples of morally inappropriate behavior often rationalized as caring by both agents and recipients include overindulgence or "spoiling," codependence, even domestic violence and incest. The care tradition may contain the conceptual resources for distinguishing appropriate from inappropriate caring but so far I have not found a convincing account (Jaggar 1995).

The ethics of care is often caricatured as a "feel good" situationist ethics that rejects justice and is concerned exclusively with personal relations; in fact most care theorists regard justice as necessary, though not sufficient, for feminist ethics; they also recognize that transforming personal relations requires transforming the larger society. Feminist ethical theory, in turn, is often equated with the ethics of care but in fact the only orthodoxy in feminist ethical theory is its broad commitment to eliminating male bias. In the next section, "Recent Directions in Feminist Ethical Theory," I indicate how this commitment has encouraged exciting theoretical developments in several ethical fields.

Recent Directions in Feminist Ethical Theory

Feminists who perceive modern ethical theory as taking over older Western-gendered binaries have several responses available to them. They may contend that women are as capable as men of realizing values culturally coded as masculine; they may embrace the hitherto devalued "feminine" pole of the binaries; they may try somehow to combine "masculine" and "feminine" values; they may deny that there is any basis for symbolizing these oppositions in gendered terms; or they may seek to rethink the conceptual dichotomies. Many liberal feminists adopt the first and/or fourth of these strategies, while care theorists tend toward the second and third; my own preferences run to the fourth and fifth. All these strategies have been used by feminists in addressing a wide variety of ethical issues.

From Practice to Theory: The Examples of Health Care, Environmental, and Development Ethics

Health care, environmental, and development ethics are often construed as fields in practical rather than theoretical ethics; however, feminist work has directly challenged the theoretical concepts that often frame discussions in each of these areas. Since the mid-1970s, the evolution of feminist health care, environmental, and development ethics has followed a trajectory that, not surprisingly, has paralleled the developments described earlier in this essay: attempts at including women's concerns have frequently been followed by charges that the theoretical frameworks are male biased; feminists then have often tried to substitute more "feminine"

frameworks, found these also problematic, and moved to various proposals for reconceptualizing each field.

Feminist health care, environmental, and development ethics each begin by noting that prevailing social practices have more severely adverse consequences for women than for men. Feminists charge that many health care practices are unjust to women, who often receive less treatment than men for the same illnesses, are allowed less autonomy, and are treated more paternalistically; for example, women's "advance directives" are more likely than men's to be disregarded. Feminists have criticized mainstream health care ethics for ignoring such injustices by failing to utilize gender in their analyses, as well as for neglecting many issues of special concern to women – or for seeing women's health issues only in terms of reproduction; they contend that health care ethics must expand its range of topics and utilize the category of gender – mediated, as always, by categories of class, race, disability, and so on (Sherwin 1992; Dula and Goering 1994; Wendell 1996). Feminists working in environmental ethics have revealed that environmental degradation often has more serious consequences for women than for men, especially for poor women and mothers, and they have argued that environmental ethics also needs to utilize the category of gender along with such related categories as caste, class, and ethnicity (Warren 1990). Similarly, feminist work in development ethics observes that "development" policies have often discriminated against women by denying them land ownership and credit; some feminists have also noted that women are disproportionately affected by the "structural adjustments" mandated by international lending institutions, which have drastically reduced the welfare functions of states in the developing world. Their conclusion, predictably, is that development ethics must also take gender into account (Moser 1993).

As in so many disciplines, feminist attempts at inclusion led to the discovery that it was impossible simply to "add women and stir" them into these areas of practical ethics; the categories available were often biased against women's experience. In health care ethics, feminists charged that prevailing conceptualizations of the "normal" patient as male – and white – led to inappropriate treatments being given to women, especially women of color, and to construing women's normal bodily experiences, such as menstruation, pregnancy, birth, lactation, and menopause, as illnesses. In environmental ethics, feminists asserted that much of environmental ethical theory was masculine: either it advocated the "mastery" of nature or, in the guise of deep ecology, it manifested a frightening disregard for real people.

Feminists have diagnosed male bias not only in development practice but also in both dominant forms of development theory – liberal modernization theory and neo-Marxist dependency theory. Both approaches cast development as a war of the sexes in which the symbolically – and often empirically – masculine must overcome the symbolically feminine – and often real women. One battle in this war is that of the public with the private; the "masculine" public sphere is portrayed as a realm of innovation regulated by universal and formally egalitarian

principles, whereas the private sphere of women and the household is portrayed as closer to nature, tradition-bound, particularistic, and stagnant. A second gendered battle is that of the nation to escape engulfment by "primordial" community and tradition and to achieve independence, self-sufficiency and self-reliance; a third battle must be waged to dominate nature and the natural. In these accounts, the nation is always masculine while community, tradition, and nature are all construed as feminine. Feminist critics deny that such gendered metaphors are simply stylistic flourishes, a form of "packaging" separable from the literal meaning of development discourse. Instead, they charge that these metaphors posit the modern (male) West as the norm of development and the traditional (female) third world as the aberration. White, bourgeois, Western men become the paradigms of maturity and progress while non-Western women are portrayed as backward, childlike, unreasonable, instinctive, and conservative, incapable of acting as agents of historical change. Man is not just superior to woman but locked in conflict with her (Scott 1996).

As alternatives to male-biased theories in health care, environmental, and development ethics, some feminists have proposed more "feminine" alternatives. For instance, they have suggested that conceptions of appropriate health care should place more emphasis on "care" as opposed to "cure"; the ethics of care has been especially influential in nursing ethics. Some ecofeminists have asserted that women have a special connection with nature; they emphasize the ethical importance of people's environmental "homes" and, in the words of one critic, portray women as caring "angels in the ecosystem." Some work in feminist development ethics seems to identify sustainable development with women's unpaid subsistence agriculture (Mies and Shiva 1993). In each of these fields, other feminists have subjected such proposals to severe criticism.

Finally, health care, environmental, and developmental ethics have developed new theoretical directions far too numerous and complex to summarize here. They include rethinking the autonomy/paternalism dilemma in medical decision making and redefining health, normalcy, illness, and reproduction in terms that are explicitly social and political as well as biological and medical. In environmental ethics, feminists are challenging ethical frameworks constructed around gendered polarities (Plumwood 1993). In development ethics, they propose to replace a preoccupation with efficiency and even welfare with a concern for women's empowerment (Sen and Grown 1987; Kabeer 1995).

Universal or Local Ethics

Rapid globalization has lent new urgency to the old question of whether ethical standards are universal. The problem has special poignancy for feminists because Western feminism has recently been preoccupied with issues of "difference," first differences between women and men, then among women; it has also been shaken by revelations of earlier Western feminism's complicity with imperialist projects.

Western feminists are seriously troubled, therefore, by the dilemma of respecting cultural difference, on the one hand, while maintaining an unwavering opposition to male dominance, on the other; in classroom discussions of this dilemma, the practice of female genital surgery has become a stock example.

One feminist response relies on the notion of "capabilities." Pioneered initially by Amartya Sen as an alternative to utilitarianism, the concept has been developed by Martha Nussbaum on whose version I focus here. Capabilities are proposed as a universal standard for measuring people's quality of life, which is to be assessed by how well a given society enables them to develop and realize their distinctively human capabilities. Nussbaum lists ten human capabilities to function, together comprising her "thick, vague conception of human nature," which she offers as a guide to development policy. She contends that this conception articulates a cross-cultural and transhistorical consensus on the central and basic human functions, reflecting "the actual self-interpretations and self-evaluations of human beings in history" (Nussbaum 1995: 72–5). However, Nussbaum offers little evidence that her list of capabilities indeed reflects a tacit universal consensus and her writings quickly dismiss disagreement; she advocates "participatory dialogue" about how postulated capabilities might be specified locally but not about which capabilities reach the list or whether list making is the best approach. My own view is that any ethical vision guiding global development must emerge from extensive and explicit democratic discussion that addresses means along with ends. For this reason, I believe that human rights, properly construed, have better credentials than capabilities as a universal standard of development.

The concept of rights was central to the emergence of Western feminism, but it is part of the same modern ethical tradition that received so much feminist criticism in the 1980s. Building on Marxist and anticolonialist critiques, some feminists contend that rights are the discourse of the dominant, so infected by their bourgeois, masculine, and Western origins that they are incapable of articulating a deep challenge either to local forms of male dominance or to a scandalously inequitable world order. Feminist charges include the following:

- Appeals to rights are often used to rationalize male power over women; for example, the right to privacy obscures domestic violence, the right to freedom of expression justifies misogynist pornography.
- Because women are not similarly situated with men, granting them formally equal rights often produces inequalities of outcome; for instance, the advent of no-fault divorce has thrown many ex-wives – but not ex-husbands – into poverty.
- Attempts to avert such outcomes by granting women "special" rights, such as maternity leave, inevitably backfire in a cultural context that conceptualizes equality as sameness. Special rights stigmatize women as inherently sexually vulnerable or as less reliable workers.
- Legal equality of rights may obscure inequalities of power to exercise them. The procedures associated with claiming and redressing rights are often

degrading, intimidating, and humiliating for women; this is especially evident in rape and sexual harassment trials.
- Women may harm themselves exercising their rights; for example, millions of women in United States alone have been harmed exercising their rights to have cosmetic surgery or to prostitute themselves. A focus on women's rights ignores the ways in which women's social situations often coerce their "choices" to exercise those rights.
- Finally, advocates of the ethics of care contend that rights talk is part of an inherently adversarial morality that disparages the more basic and important human values of interdependence, cooperation, and trust.

For all these reasons, feminist critics charge that not only may rights talk be unhelpful to women but it may even rationalize inequality.

Other feminists, including myself, believe that the rights tradition has the conceptual resources to address those charges. For instance, rights may be interpreted to take account of morally salient differences among rights holders and they may be assigned to groups as well as individuals. They may also include "positive" as well as "negative" rights; these are "entitlements" rather than liberties and they carry claims not only to noninterference but also to correlative duties on the part of others. Such rights may be thought of as embodying the values of community, mutual aid, and social solidarity.

Those who regard rights as indispensable for women's liberation look to the burgeoning global feminist movement inspired by the slogan, "Women's rights are human rights." The theorists of this movement have followed the now-familiar evolution of much feminist ethics; beginning by criticizing abuses of women unrecognized as rights violations, they moved to challenging the covert male norm concealed in traditional conceptions of so-called human rights and to proposing radical reinterpretations. Space does not permit a full account of these feminist proposals but central to them is a recognition that violations of women's human rights are typically carried out by nonstate as well as state actors – often by male family members – and that they occur in the private as well as the public sphere. This recognition requires expanding the definition of state sanctioned repression to include acceptance of family forms in which brides are sold and in which fathers and husbands exert strict control over women's sexuality, dress, speech, and movement. Slavery must be defined to include forced domestic labor and prostitution. Because some violations of human rights are gender specific, the definitions of war crimes and genocide must be expanded to include systematic rape, sexual torture, female infanticide, the systematic withholding of food, medical care and education from girls, and the battery, starvation, mutilation, and even murder of women. Feminists have also highlighted the link between violations of women's civic and political rights and violations of their economic and social rights: economies as well as laws dictate the worldwide preference for boys over girls and women's economic vulnerability exposes them to the more blatant abuses (Peters and Wolper 1995).

The male bias and consequent false humanism of older conceptions of human rights should be corrected, perhaps by imagining the normative human as female rather than male. Women, after all, are vastly overrepresented among the poor and illiterate of the world and they are certainly those most vulnerable to oppressive systems of power.

The slogan "Women's rights as human rights" emerged from a grassroots activist movement. It avoids metaethical questions regarding the grounding of rights, a concept that Bentham notoriously characterized as nonsense on stilts. Instead, the slogan indicates a vision that is ethical rather than metaphysical, a vision that has a good claim to expressing the "overlapping consensus" of feminists worldwide. As will become evident in the next section, "Rethinking Ethical Theory," I believe that this is the only justification it needs.

Rethinking Ethical Theory

The term "ethical theory" covers a wide range of intellectual enquiries, all of which share an interest in questions about morality in general rather than in immediately practical ethical concerns. The questions addressed by ethical theory range from the metaphysical, such as whether there exists an objective moral realm, to the epistemological, most notably, how moral claims may be justified, to more directly normative but still general inquiries into such central ethical notions as the good, the right, and the just. For much of the twentieth century, Western ethics was dominated by a particular understanding of ethical theory but, as one century has drawn to a close and another one commenced, this model has been increasingly challenged. Some feminists are among its most outspoken challengers.

The idea of ethical theory now brought into question is one that Margaret Walker labels the theoretical-juridical model. Walker traces this conception of ethics back to Henry Sidgwick, noting that his career as Knightsbridge Professor of Moral Philosophy at Cambridge (1883–1900) marked a decisive shift toward academic specialization and the professionalization of philosophy and other "disciplines" in the universities. According to Sidgwick, the job of ethics or the philosophical study of morality was not to determine the right or reasonable thing to do in particular situations but rather "to seek systematic and precise general knowledge of what is *right* and what makes judgments *valid*." Sidgwick regarded this project as unlike science in that its task was to formulate regulatory rather than explanatory laws but as scientific in form because it sought systematic, precise, and general knowledge through a method guided by disinterested demands of precision, clarity, and consistency. Thus Sidgwick defined what Walker calls "the idea of a *pure core of knowledge* at the heart of morality," a core that excluded both empirical contributions from the social sciences and considerations of the

historical and cultural placement of "our" moral views (Walker 1998: 35). He fathered the idea of an ethical theory as

> a consistent (and usually very compact) set of law-like moral principles or procedures for decision that is intended to yield by deduction or instantiation (with the support of adequate collateral information) some determinate judgment for an agent in a given situation about what it is right, or at least morally justifiable, to do.
>
> (Walker 1998: 36)

In the philosophical tradition stemming from Sidgwick, the aim of ethics is to discover/construct, test, compare, and refine ethical theories, and an individual's moral capacity is pictured as a kind of theory within "him." Ethics is juridical both in constructing theories that deliver verdicts on particular practical problems and in adjudicating theories for their (logical and epistemological) adequacy (Walker 1998: 37).

Sidgwick's definition of ethics makes it clear that he regarded justification as central to ethical theory – though later in his career he changed this view. Sidgwick's successors maintained his epistemological focus but developed a distinct approach to addressing his questions. G.E. Moore's *Principia Ethica*, published in 1903, is credited with initiating the linguistic turn in ethical theory, directing philosophical attention away from explicit consideration of normative issues and refocusing it on the language and logic of ethics. Moore and his heirs construed their project of analyzing ethical language and logic in terms reminiscent of Sidgwick; they pursued systematic and general knowledge of moral concepts through a careful, clear, and disinterested inquiry into uses of moral language. Ignoring as irrelevant most social facts, including facts about the social situations of those who used the language they analyzed, they sought to discover universal ethical truths, apparently assuming that the concepts implicit in "our" moral language were transhistorical and transcultural. Although "metaethical" analyses were expected to be impartial and dispassionate, many philosophers hoped that they could generate substantive theoretical conclusions. Moore believed that determining the meaning of "good" would reveal what was intrinsically good; R.M. Hare claimed that his analysis of "*The Language of Morals*" (1952) excavated the fundamental principles of right.

Although the term "applied ethics" is still often used to refer to thinking about practical moral questions, the second half of the twentieth century and the early years of the twenty-first century have seen a steady erosion of the idea of ethical theory as a product of (more or less) pure reason to be "applied" to practical issues. The hold of this deductivist model has weakened in even the early work of John Rawls, whom many regard as the twentieth century's ethical theorist *par excellence*. Rawls's concept of reflective equilibrium, reached by weighing general principles and considered moral judgments against each other with an openness to modifying either, exemplifies a model of moral reasoning that does not privilege

theory over intuition. In his more recent work, Rawls lowers even further the status of ethical theory by abandoning the ideal of an "Archimedean point" and substituting the notion of a specific community's "overlapping consensus" about justice. Other philosophers have developed additional challenges to the Sidgwickian conception of ethical theory. Walker observes that, "Code-like theory has provoked criticism from Aristotelians, Humeans, communitarians, contemporary casuists, pragmatists, historicists, Wittegensteinians, and others in the last several decades. So clear is this schism in late-twentieth-century moral philosophy that talk of 'antitheory' in ethics is now familiar" (Walker 1998: 53–4). One way in which some feminists have challenged this model is by contending, as do some care theorists, that ethical theory is entirely dispensable as guide to right action.

A few feminists have added a moral and political dimension to the increasingly common claim that the codelike conception of ethical theory offers a misleading model of moral justification. Margaret Walker charges that this conception conceals the specific, partial, and situated character of views and positions that are put forward "authoritatively as truths about 'human' interest, 'our' intuitions, 'rational' behavior, or 'the' moral agent" (Walker 1998: 54). The cloak of scientific objectivity woven by Sidgwick "signifies the promise and ensuing prestige of scientific accomplishment," while shielding from view "the historical, cultural, and social location of the moral philosopher, and of moral philosophy itself, as a practice of authority sustained by particular institutions and arrangements" (Walker 1998: 56).

The accounts of moral justification predominant in modern ethical theory invoke such ideals as rationality, universality, impersonality, detachment, dispassion, neutrality, and transcendence. They aspire to evaluate actions and practices from a postulated "moral point of view" often explicated in metaphorical terms such as a god's eye view, the perspective of an ideal observer or an archangel, an Archimedean point, a view from nowhere, or a view from everywhere. These metaphors are paradoxical, of course, since their aim is to designate an imagined perspective that is precisely not a specific point of view.

A number of feminists have argued that philosophers' claims to articulate the moral point of view have often in fact described only the view from their chairs in the gentleman's club; however, the particularity of their views has been concealed by the falsely universal pretensions implicit in their accounts of their methods for justifying them. Elizabeth Anderson contends that G.E. Moore's account of "our" moral intuitions in fact reflected the beliefs of those with the most social power even in Moore's narrow, elite – and overwhelmingly male – circle, and she suggests that this biased outcome was not accidental but reflected a tendency endemic to individualist intuitionism (Anderson 1993). Elsewhere, I have traced similar feminist challenges to the accounts of justification produced by other prominent twentieth-century philosophers (Jaggar 2000). One common theme linking those challenges is feminists' claim that people's perceptions, values, and modes of reasoning, their understanding of their own and others' needs and interests, even their constructions of moral situations, vary not only individually but also systematically according to their social experience and locations. For this

reason, feminist critics charge, the recommendations characteristic of many modern accounts of justification – recommendations that moral agents should put themselves in the place of others, think from others' perspectives, reverse perspectives with them, and so on – are epistemically incoherent. Although such thought experiments may have rough-and-ready heuristic value, there is no reason to suppose that philosophers' imaginations are more insightful or reliable than anybody else's. Claiming that their imagination enables them to fly up to attain "the moral point of view" is a disingenuous rhetorical device by which some philosophers lend a tone of magisterial authority to their own pronouncements.

By refusing to acknowledge the effects of people's social identities on their moral understandings, the ideal of point-of-viewlessness insulates itself from any critical examination of its own social origins or functions. Specifically, it denies that any philosophical significance attaches to the fact that only a few persons are authorized to define moral knowledge. Yet, as Walker notes, "To have the social, intellectual, or moral authority to perform this feat, one must already be on the advantaged side of practices that distribute power, privilege, and responsibilities in the community in which one does it" (Walker 1998: 54). From "the" moral point of view, the fact that "Western Anglo-European philosophical ethics as a cultural tradition and product has been until just recently almost entirely a product of some men's – and almost no women's – thinking" is a matter of only historical, not philosophical, interest. Some feminist critics charge that traditional conceptions of the moral point of view are more than expressions of a juridical-administrative perspective that some have characterized as masculine; nor have they served only as a means of rationalizing practices oppressive to women and members of other subordinated groups; they have even had a function beyond that of invalidating criticism of these practices. Invoking them has been, finally, a means by which philosophers have rationalized their claims to define the criteria of ethical justification and to judge when those criteria have been met.

A number of feminist philosophers, myself included, are working to rethink how moral justification might be more transparent and less covertly elitist and authoritarian. Like Habermas, we see empirical discourse as indispensable to justifying particular claims in particular contexts; however, we find that Habermas's account has only limited usefulness, in part because his ideal of "domination free" discourse imposes such stringent and counterfactual conditions that it seems virtually impossible for it to be achieved in real life. In contrast with the idealizations characteristic of most philosophical accounts, we address directly the inevitabilities of cultural difference and hierarchy among participants in empirical discourses. Although these new feminist understandings of moral justification are less idealized and more naturalistic than traditional accounts in that they operate at a lower level of abstraction, they are not naturalized in the classic Quinean sense of claiming "scientific" or value-neutral status. On the contrary, they are still explicitly normative, linking the development of increasing moral objectivity with developing justificatory practices that are increasingly egalitarian, democratic, and inclusive (Benhabib 1992; Jaggar *et al.* 1995; Walker 1998).

Feminist challenges to mainstream ethical theory have pursued transparency by making visible what Walker calls "the gendered structures of authority that produce and circulate (moral) understandings" (Walker 1998: 73). However, such challenges are far from constituting a wholesale rejection of ethical theory, even modern ethical theory. As Walker observes, feminist demands for transparency are embarrassing precisely because they invoke the familiar modern values of representation, consent, self-determination, and respect. They appeal to values that "are of specifically democratic, participatory, and egalitarian kinds, squarely founded on moral and political ideals of modern Western social thought" (Walker 1998: 73).

Contemporary feminist philosophers continue to do ethical theory in the sense of thinking generally about morality and, like other ethical theorists, we lean heavily on the analysis of language. However, our analyses differ from those of mainstream ethical theory in several respects. For one thing, we examine more than the strictly logical aspects of language: we look at metaphor, symbol, and "nonlogical" implications (Calhoun 1988), including emphases, omissions, and silences, and we pay attention to the moral and political significance of these aspects. We often say we are analyzing "discourses" rather than "language," thereby suggesting that we are examining conceptual frameworks that are multiple, contingent, and disclose as much about the authors of the discourse as they reveal about the moral realities to which ethical discourses purport to refer. We question the implicit normativity of terms like "we," "our," and "ordinary language." We take as objects of our scrutiny the discourses of philosophy; these include not only the discourses of our colleagues, which construct what is authorized and taught as ethical theory, but also our own discourses, which often construct what is authorized and taught as feminist ethical theory.

Our version of feminist ethical theory thus has several distinguishing features. First, it utilizes the categories of gender and other inseparable categories of social difference and hierarchy on the levels of theoretical as well as practical ethics. Second, it enlarges the domain of ethics to include ethics itself: we undertake the ethical analysis of ethical analysis, the ethical theory of ethical theory. We see contemporary ethical theory as a discourse situated in a larger society and we ask who defines it and how their – and its – authority is maintained; we also see ethical theory as a professional practice and so we are led to examine such aspects of that practice as canon formation, prizes, prestigious offices, and lectureships. Finally, our work is – or aspires to be – distinguished by its self-reflectiveness: we try to be conscious of the assumptions and implications of our own ethical theorizing, including their practical consequences; we seek to produce ethical theory that we acknowledge to be partial and provisional from our own explicitly situated perspectives.

References

Anderson, Elizabeth (1993) *Value in Ethics and Economics*, Cambridge and London: Harvard University Press.

Baier, Annette C. (1987) "The Need for More than Justice," in *Science, Morality and Feminist Theory*, eds. Marsha Hanen and Kai Nielsen, Calgary, Canada: University of Calgary Press.

Benhabib, Seyla (1992) *Situating the Self: Gender, Community and Postmodernism in Contemporary Ethics*, New York: Routledge.

Blum, Lawrence A. (1982) "Kant's and Hegel's Moral Rationalism: A Feminist Perspective," *Canadian Journal of Philosophy* 12 (2): 287–302.

Blum, Lawrence A. (1992) "Care," in *Encyclopaedia of Ethics*, ed. Lawrence C. Becker, New York: Garland.

Calhoun, Cheshire (1988) "Justice, Care, Gender Bias," *Journal of Philosophy* 85 (9): 451–63.

Card, Claudia (1990) "Gender and Moral Luck," in *Identity, Character and Morality*, eds. Owen Flanagan and Amelie Rorty, Cambridge: MIT Press, pp. 199–218.

Clark, Lorenne M.G. and Lange, Lynda, eds. (1979) *The Sexism of Social and Political Theory*, Toronto, Buffalo, London: University of Toronto Press.

Dula, Annette and Goering, Sarah (1994) *It Just Ain't Fair: The Ethics of Health Care for African Americans*, Westport and London: Praeger.

Friedman, Marilyn (1993) *What Are Friends For? Feminist Perspectives on Personal Relationships and Moral Theory*, Ithaca: Cornell University Press.

Gilligan, Carol (1982) *In a Different Voice: Psychological Theory and Women's Development*, Cambridge, MA: Harvard University Press.

Held, Virginia (1993) *Feminist Morality: Transforming Culture, Society, and Politics*, Chicago: University of Chicago Press.

Hoagland, Sarah Lucia (1988) *Lesbian Ethics: Toward New Value*, Palo Alto, CA: Institute of Lesbian Studies.

Jaggar, Alison M. (1995) "Caring as a Feminist Practice of Moral Reason," in *Justice and Care: Essential Readings in Feminist Ethics*, ed. Virginia Held, Boulder: Westview, pp. 179–202.

Jaggar, Alison M. (2000) "Feminism and Moral Justification," in *The Cambridge Companion to Feminism in Philosophy*, eds. Miranda Fricker and Jennifer Hornsby, Cambridge University Press, pp. 225–44.

Jaggar, Alison M., Sterba, James P., Fisk, Milton, *et al.* (1995) *Morality and Social Justice: Point Counterpoint*, Lanham, MD and London, UK: Rowman & Littlefield.

Kabeer, Naila (1995) *Reversed Realities: Gender Hierarchies in Development Thought*, New York: Verso.

Kohlberg, Lawrence (1981) *The Philosophy of Moral Development: Moral Stages and the Idea of Justice*, San Francisco: Harper & Row.

Mies, Maria and Shiva, Vandana (1993) *Ecofeminism*, London: Zed Press.

Moser, Carolyn (1993) *Gender, Planning and Development*, London: Routledge.

Noddings, Nel (1984) *Caring: A Feminine Approach to Ethics and Moral Education*, Berkeley: University of California Press.

Noddings, Nel (1990) "Feminist Fears in Ethics," *Journal of Social Philosophy* 21(2–3): 25–33.

Nussbaum, Martha (1995) "Human Capabilities, Female Human Beings," in *Women, Culture and Development,* ed. Martha Nussbaum and Jonathan Glover, Oxford: Clarendon Press, pp. 61–104.

Okin, Susan Moller (1979) *Women in Western Political Thought*, Princeton: Princeton University Press.

Okin, Susan Moller (1989) *Justice, Gender and the Family*, New York: Basic Books.
Peters, Julie and Wolper, Andrea, eds. (1995) *Women's Rights, Human Rights: International Feminist Perspectives*, New York: Routledge.
Plumwood, Val (1993) *Feminism and the Mastery of Nature*, London: Routledge.
Ruddick, Sara (1989) *Maternal Thinking: Towards a Politics of Peace*, New York: Beacon Press.
Scott, Catherine V. (1996) *Gender and Development: Rethinking Modernization and Dependency Theory*, Boulder: Lynne Rienner.
Sen, Gita and Grown, Caren (1987) *Development, Crises and Alternative Visions: Third World Women's Perspectives*, New York: Monthly Review Press.
Sherwin, Susan (1987) "A Feminist Approach to Ethics," *Resources for Feminist Research* 16 (3).
Sherwin, Susan (1992) *No Longer Patient: Feminist Ethics and Health Care*, Philadelphia: Temple University Press.
Tronto, Joan C. (1993) *Moral Boundaries: A Political Argument for an Ethic of Care*, New York: Routledge.
Walker, Margaret (1998) *Moral Understandings: A Feminist Study in Ethics*, New York: Routledge.
Warren, K.J. (1990) "The Power and Promise of Ecological Feminism," *Environmental Ethics* 12 (2): 125–46.
Wendell, Susan (1996) *The Rejected Body: Feminist Philosophical Reflections on Disability*, New York: Routledge.
Whitbeck, Caroline (1984) "A Different Reality: Feminist Ontology," in *Beyond Domination*, ed. Carol Gould, Totowa, NJ: Rowman & Allanheld, pp. 64–88.
Young, Iris Marion (1990) *Justice and the Politics of Difference*, Princeton: Princeton University Press.

Chapter 21

Continental Ethics

William R. Schroeder

Introduction

This essay will explore some key features of Continental ethics through an examination of several especially creative ethical thinkers in the Continental tradition: Hegel, Nietzsche, Scheler, Sartre, and Levinas. After an introductory discussion of relationships between Continental and analytic approaches to ethics, I describe some of the major innovations of each of these philosophers. Then I examine some apparent disagreements among them and conclude by describing some of their challenges to mainstream ethical theory. My goal is to provide a sense of the originality and diversity of ethical positions among Continental thinkers.

Three features of Continental ethics can provide benchmarks to establish the contours of the landscape. The first feature is that Continental ethicists have blazed some of the trails that are producing a virtual revolution in current ethical theory in the analytic tradition. The second feature is a common suspicion among Continental thinkers that seems almost entirely absent from analytic ethical theory, at least until quite recently, namely, that there is something deeply suspect – even immoral – about morality itself. The third feature is a common project among Continental ethicists that establishes a different tone than that typically found in analytic ethical theory; this project is finding the conditions that will enable genuine personal flourishing – or ethical radiance – among individuals. Far less emphasis is given to duty and obligation in Continental ethics and far more attention is paid to the cultural, psychological, interpersonal, and emotional conditions of a personal transformation that makes serious ethical achievement possible.

The Blackwell Guide to Ethical Theory, Second Edition. Edited by Hugh LaFollette and Ingmar Persson.

The first feature concerns the many areas of common interest between Continental thinkers and recent analytic ethicists. In the past thirty years, there has been a two-pronged revolution in analytic ethics. One prong involves a search for alternatives to the long-reigning hegemony of Kantian deontology and utilitarian consequentialism. In the process a variety of major assumptions that have long guided ethical thinking are being challenged. Perhaps the most notable emerging alternative is the renewal of virtue ethics, which centers ethical practice on the formation of character dispositions and clarifies the main components of the virtues and the ways they might be cultivated (Slote 1992; McIntyre 1981; Foot 1978; Wallace 1978). I would argue that Nietzsche began this project (in its contemporary form) and that Nicolai Hartmann makes central, if little known, contributions to it. Another important development is a renewed interest in moral realism as a way of restoring some degree of objectivity to ethical aspirations and assessments (Murdoch 1970; McDowell 1985; Boyd 1988). Max Scheler developed an intriguing version of value realism that might offer some useful handholds to help scale this mountain. Relevant too are efforts to recuperate the history of ethics by clarifying goods dominant in previous eras and thus to deepen our responsiveness to their continuing resonance in this era (Taylor 1989). This approach to all fields of philosophy – not just ethics – was pioneered by Hegel who made important contributions. Feminists are challenging the duty-based, rationalist orientation of some ethical theories, suggesting that they ignore the taproot of ethical action – sensitivity and nurturing – and miss a more particularized grasp of lived situations that can navigate through ethical complexities without abstractions or principles (Nussbaum 1990; Noddings 1986). Scheler and Nietzsche made some important contributions to this project as well. Another important development questions whether all positive values are comparable and commensurable and whether they can be integrated into a coherent system (Nagel 1979; Stocker 1990). Nietzsche, Scheler, and Sartre stress this potentially tragic fact about fundamental values.

This diverse array of new experimental directions in analytic ethics is supplemented by a second, equally important, development – the vigorous growth of applied ethics. Applied ethics attempts to refine commonly accepted moral intuitions to make them more coherent, and to utilize the socially established purposes of specific institutions to provide a rational foundation for normative guidelines for actors within such institutions. Applied ethics takes the basic functions of an institution for granted; it resolves tensions and hard cases by ordering these functions. Such social institutions constitute what Hegel called the ethical substance of a society – established traditions and procedures for achieving commonly valued outcomes. Each institution (e.g., medicine, law, business, engineering, academia) functions as a limited frame within which genuine resolution of major disputes might be achieved. Hegel's *Philosophy of Right* might be considered the first treatise in applied ethics – first addressing the sphere of the family, then the sphere of civil society, and finally that of the State. He does not so much try to justify the values

within these spheres as render them coherent. He searches for the rational elements of those institutions and attempts to discover their most effective realization. My contention is that both prongs of these recent developments in analytic ethics might benefit from some understanding of Continental contributions. The brief expository sketches in the next section, "Some Major Figures," offer some evidence to support this claim. For the moment I have only indicated common directions.

The second feature of Continental ethics sharply distinguishes it from much of current analytic ethics, namely its ethically motivated suspicion of conventional morality and a skepticism about the value of many commonly accepted moral values. Continental figures find conventional morality deficient, even ethically objectionable. They also regard much traditional moral theory to be misguided and bankrupt. In addition, they are skeptical about the psychology and philosophical anthropology that informs much of everyday morality and ethical theory, and they try to discover alternatives. Nietzsche's tirades against Christian morality and its Enlightenment heirs offer only the most flamboyant example of this line of thought. Additional examples include Hegel's attacks on the emptiness of Kantian ethics, Scheler's attacks on common mistakes of rationalism in philosophical ethics (and their ruinous ethical results), Sartre's reservations about the world of the Good (and its consequent need for Evil), and Levinas's doubts about the interest-seeking character of politics. Analytic thinkers want to believe the house of ethics is basically shipshape, needing only a coat of paint and a good cleaning, but Continental figures think the traditional house has become a prison and new habitats must be created.

The third feature follows from the second. It consists in an entirely different orientation among Continental thinkers. Their primary goal is to determine the conditions of ethical flourishing for various types of individuals, rather than to establish the borders of behavior that must not be transgressed. Since they have no confidence in traditional frameworks and guidelines, they seek to discover new paths that will facilitate radiant ethical achievement. Again, this is perhaps most evident in Nietzsche, who spends even more time exploring the psychological, physiological, and cultural conditions of becoming a "greater" person (more enriched, more varied, subtler, and stronger) than he does on his genealogical critiques of modern values. But this is also true of Hegel, who attempts to discern the social institutions that will most support the achievement of individual harmony and recognition. And of Scheler, whose examinations of the diverse strata of value and of the role of feeling and basic moral tenor in ethical response reveal ways to strengthen one's strongest ethical capacities. And of Sartre, who explores an experience he calls "conversion," which he regards as a precondition of personal authenticity. And of Levinas, who believes that a special orientation toward other people is a necessary precondition for ethical life. For Continental figures, one function of ethical theory is to elucidate a transformation of personal existence that enables a more promising kind of ethical practice.

Some Major Figures

G.W.F. Hegel (1770–1831)

Hegel reacted against the formalism (lack of substantive norms) and the abstractness of Kant's ethics. He rejects the idea that one can test ethical maxims on the basis of any single formula (e.g., the Categorical Imperative), that an action can be genuinely motivated solely by duty, and that individuals can live ethically in isolation from rational social contexts and institutional norms. Hegel shows that the Categorical Imperative (the demand that maxims be universalizable without contradiction) is ultimately contentless – sanctioning almost any action, depending on how the "maxim" of the action is interpreted and which features of the situation are deemed relevant (Hegel 1807/1977: 256–62). He claims that adopting the moral viewpoint may be ethically counterproductive because it forces one to abstract oneself from all the concrete social ties (one's relation to one's family and profession) which give ethics its point and structure (Hegel 1807/1977: 365–83). Hegel is profoundly skeptical of the whole project of generating moral precepts a priori – on the basis of pure reason alone. Instead, he insists that individual flourishing is encouraged by social institutions, that the direction of individual self-realization is related to the goals of those social institutions that surround them, and that these structures provide the basic normative content of ethics.

In his earliest writings Hegel also reacts against the experience of division and conflict typical of modern social life. Individuals experience sharp conflicts between reason and passions (a division only exacerbated by Kant's uncompromising rigorism about moral motivation); they also often experience themselves opposed to other individuals and to the larger political order. One of Hegel's main aims is to produce a theory that might overcome these divisions. Overcoming this divided and alienated condition is required for achieving genuine freedom, and Hegel takes freedom thus understood to be the defining goal of human life. Such harmonious freedom is a humanly created achievement (as opposed to a naive unity that is easily lost). The right institutions are needed to make this harmony possible. Reason is one important means by which these mediated unities are produced. Hegel realizes that his conception of freedom is not the same as the negative freedom sought throughout the modern world – removing obstacles to doing what one wants. That freedom dominates modern economic life. Hegel does not reject this characteristically modern aspiration, but asserts that the proper social institutions are needed to fulfill other human aspirations and insists that social institutions can enhance as well as limit human choices. Hegelian individuals do not see themselves as isolated from or imprisoned by his ideal social institutions, but instead see themselves expressed and realized in them (Hegel 1821/1967: 105–10).

Reciprocal recognition is the way Hegelian freedom is realized in interpersonal life. It involves reciprocal acknowledgment that others one encounters are fully

self-conscious like oneself. The process of achieving this acknowledgment also enriches the self-consciousness of each person. Prior to social acknowledgment, self-consciousness is only nascent; after the process of acknowledgment the identity of each participant as both living and self-conscious is ratified. In becoming socially recognized one becomes more actual for oneself – more fully self-conscious (Hegel 1807/1977: 111–13). Reciprocal recognition transcends a stage of encounter in which individuals violently struggle to prove their superiority. The outcomes of these battles leave at least one party unrecognized and dominated by the other. Though such domination initially seems to achieve the aspirations of self-consciousness, it does not really do so (the slave's recognition of the master is corrupted by the fact that he is a slave, and the slave's labor leaves the master's talents underdeveloped) (Hegel 1807/1977: 115–19). When two persons achieve reciprocal recognition, they experience a deeper harmony; they no longer see each other as alien. Both produce the harmony together, and both are enriched by it. It establishes a genuine community. To be recognized in one's performance of a social role is to achieve social actuality.

Hegel suggests that there are three spheres in which persons can experience themselves as realized/expressed (or, alternatively, as foreign or alienated): in relation to things, to their own willed actions, and to communal life generally. A sense of connection with material goods can be achieved through appropriate property relations; a material thing becomes one's own by becoming one's property. Property is thus essential to individual freedom for Hegel (Hegel 1821/1967: 40–4). But property relations cannot be established in isolation. Complex legal understandings and a judicial enforcement mechanism are necessary to sustain property relations. Even at the lowest levels of freedom's realization, an entire institutional system is presupposed. So important is property that Hegel suggests that the rational State will make some modest amount of property available to everyone to ensure this modest degree of self-actualization (Hegel 1821/1967: 45, 57–64).

The second sphere is one's own practical activity. Here an action is one's own only if it derives from one's free decision. Hegel accepts Kant's claim that action is not truly one's own if one follows the commands of others or if one unthinkingly conforms to tradition. But he does not go so far as to claim that freedom requires performing the act solely for duty's sake alone. Any practical action expresses one's joy in exercising one's capacities; this can never be purged from action without disabling it altogether. Moreover, reason alone will not supply any substantive goods or duties. These derive from the concrete communities in which one lives – as a member of a family, a profession, and a state (Hegel 1821/1967: 105–10). Duties, goods, and norms derive from these concrete relations. They are the substance of ethics, but they will be expressions of oneself as a rational being only if they exhibit a coherent order.

Thus, neither the first nor second spheres of ethical life can stand alone; specific social and institutional relations are necessary to achieve full freedom in them. The state and other social institutions play a central role in Hegel's theory because they make individual freedom possible. If institutions are irrationally organized or

inconsistent, then they will undermine the achievement of freedom. Rational institutions are necessary if there is to be any freedom – any sense of social harmony. This is the source of the positive role of the social institutions (including the state) in the self-actualization of individuals. By breathing life into their social roles, individuals make the social order express themselves, and by ensuring that these roles are coherent and conducive to self-realization, social institutions enhance the identity of individuals (Hegel 1807/1977: 266–78).

But not just any social and political order is rational. A rational social order must satisfy various criteria (Hegel 1821/1967: 155–60, 174–6): It must guarantee property relations and ensure that each person owns at least some property. It must facilitate interpersonal recognition and cultivate the capacity for moral reflection and deliberation. It must provide institutional means for mastering a profession (allowing individuals to create their own self-identity) and participating as a citizen. It must regulate the economic sphere in which individuals pursue their own desires, and it must respect the integrity of individuals. It must prevent conflict among professions and mediate disagreements over property and rights. It must also not allow citizens to fall into the illusion of isolated individuality, abdicating their responsibilities to the state and other social institutions. The state (and all the mediating institutions of society, including families and guilds) thus plays an active and essential role in making freedom possible for Hegel, and there are many ways for existing states to fall short of these evaluative criteria. For Hegel, institutions and political organization are essential to ethical life because they provide the proper environment for achieving expressive, integrated lives.

Friedrich Nietzsche (1844–1900)

Nietzsche is probably the most sweeping critic of traditional moral ideals and moral philosophy – far-reaching enough to call himself an "immoralist." Fortunately, he does not only demolish; he also constructs an alternative orientation to, and a different basis for, ethics. Nietzsche's major nemesis is Christianity and its modern offshoots. Among these offshoots he includes many Enlightenment assumptions and ideals (liberty, equality, fraternity). Nietzsche sensed the imminent collapse of religion as a serious intellectual option and correctly predicted a resulting rise of nihilism. He sought above all to avoid the worst effects of this reaction (Nietzsche 1882/1974: §125). To overcome nihilism humanity must transcend the fantasies of religion and commence a new approach to life. Nietzsche's entire project is to enable the transition to this new way of living. He offers a series of alternative ideals to guide the process and produces a more penetrating human psychology in order to achieve realistic progress. He also describes some cultural changes that may contribute to this transition.

Nietzsche attacks both the moral point of view and traditional moral values, especially Christian values. By "morality" he means a system of duties and

obligations ("shalts") that present themselves as universally justified imperatives. He issues two challenges to morality in general: first, a challenge to its direction; and second, a demonstration of its immoral sources. Together these objections establish the need for a new alternative.

Nietzsche offers a number of arguments to show the general misguidedness of morality. First, he clarifies the basic effect of morality on those who try to live up to it. Through guilt, morality typically turns people against themselves, undermining their sense of passion, self-esteem, and creativity; often it makes them cower before some external principle or power. It makes people weak, fearful, ashamed of themselves, self-effacing, and dependent (Nietzsche 1881/1982: §18; 1887/1968: III §11). It does this in part because it uses fear as its central motivation and in part because it extirpates and eradicates the very foundations of any organism and of life itself: sexuality, self-expression, the exercise of power (Nietzsche 1888/1954: 486–92). Nietzsche seeks to create a different ambiance: a sense of innocence, of new birth, an experimental light-footed approach to life, a stimulus to continually reach beyond one's current achievements (Nietzsche 1883–5/1954: I §1, I §22). Though it may take considerable effort to achieve this stance, it embodies Nietzsche's goal of living "beyond good and evil."

Most moral codes are repressive and puritanical; they make impossible demands on people, and the failure to live up to these demands produces people who hate life and themselves (Nietzsche 1883–5/1954: I §3). Even if one succeeds in being moral, one is rarely a radiant, joyful person. Instead one often seethes with repressed aggression and resentment. Instead, the approach Nietzsche would substitute works to cultivate the basic instincts of life and humanity, to increase humanity's capacity for self-transformation, and to make possible a general affirmation of life (despite its sufferings, accidents, and "amorality"). He tries to educate people to create goals for themselves, to live with buoyancy, grace, and confidence, finding joy in their actual attainments (Nietzsche 1883–5/1954: IV §13). None of this, he contends, is typically associated with the moral point of view or with being moral.

In addition, morality typically demands automatistic, habituated responses – abiding by tradition (or duty) for tradition's (or duty's) sake (Nietzsche 1881/1982: §19). Nietzsche thinks this makes serious, supple responses to the complexity of actual situations almost impossible. Nietzsche's formula for this effect is that morality turns people into herd animals, thus inhibiting subtle responses to complex situations. Also, most moralities stress one's duties to other people (vs self-perfection) and the urge to judge others (or oneself) rather than achieve self-transformation. Finally, morality often arrogates itself to the mantle of the sole hegemonic standard for evaluating people, and thus devalues many other criteria on which they can be assessed: physical, intellectual, emotional, vital, artistic, interpersonal, and political. Moral excellence is only one dimension of human achievement, but morality pretends it is the only important – or at least the most significant – measure of any consequence. Nietzsche simply rejects this; serious personal evaluation would include all these dimensions. Nietzsche himself adds

the following tests: health and vigor, the capacity to affirm life, strength and hardness, and the amount of truth one can bear (Nietzsche 1886/1968: §227). Morality is a minimal standard at best; Nietzsche expects more of people, and advocates self-perfection along many dimensions.

Nietzsche's second general line of attack is a series of points designed to show that morality is rooted in or dependent upon immoral sources or at least wholly nonmoral standpoints. First among these is the claim that the decision to be moral cannot itself be a moral decision and cannot be defended on moral grounds on pain of circularity (Nietzsche 1881/1982: §97). Also, many motivations for behaving morally are not themselves very admirable, for example, fear, despair, egoism, habit, fanaticism. Similarly the standpoint from which one must evaluate morality as a cultural-social institution cannot itself derive from the morality being evaluated. Any judge must somehow achieve an extra-moral standpoint – overcoming the biases of years of moral indoctrination to be able to explore the many perspectives needed to provide a fair evaluation of diverse moral codes. Still another point is that moral innovators almost always are called "immoral" by the then dominant morality, and hence the dominant morality subverts efforts to improve the institution as a whole.

An entirely different type of challenge derives from Nietzsche's examination of Christian morality, namely, that many historical moralities to date have been "slave" moralities. Slave moralities develop by undermining and transvaluing "master" moralities (Nietzsche 1886/1968: §268). Nietzsche suggests that "master moralities" are those in which flourishing human beings bestow value on the traits that make their flourishing possible. Everything that produces delight, joy, power – everything that affirms this life – is sanctified with the appellation "good." For such moralities "bad" is almost an afterthought – indicating simply a lack of good traits; the bad person is less condemned than unfortunate. However, those who are more impoverished, less advantaged, and less talented resent and hate the success of those who flourish. They seethe with unbounded hatred. To avenge themselves, they create the category of "evil." They substitute for the opposition "good/bad" the more vituperative opposition of "good/evil," and they invert what counts as good (Nietzsche 1887/1968: I §5–10). How else could meekness, self-effacement, and poverty have become virtues? Thus, for slave moralities "evil" becomes whatever master moralities have called good – whatever has made human flourishing possible. And "good" becomes whatever was previously merely bad – characteristics indicating misfortune, weakness, disability, self-denial, and life-denial. Thus Nietzsche accuses morality of inverting what is truly valuable and exhorting humanity in the wrong direction. It has transformed what is truly good into something evil, and turned mediocrity into virtue. Thus it inhibits and destroys the best and most promising traits in the human species.

Nietzsche inaugurates a revaluation of values, in which specific values inherited from the past are tested and assessed to determine their overall promise. Nietzsche explores the actual effects of living in accord with specific ideals for specific human types. He investigates a number of types of value (epistemic, aesthetic, moral,

religious) as well as important modern values, for example, equality, happiness, democracy, truth, and many Christian values, for example, love, compassion, and humility. His method of examination involves interrogating a value from many different points of view (Nietzsche 1888/1954: Preface): What is the typical result of embodying the value? What are the typical motives for pursuing the value? What is the manner in which the value is typically lived? What alternatives are there to this value? Who typically pursues this ideal and how does it function in practice? By answering these questions Nietzsche is able to challenge a wide variety of traditional values. Often his investigations show how the value might be refurbished or transfigured. His aim is not simply to reject or debunk, but to discover which values still have something to offer the future, life-affirming development of humanity.

For example, Nietzsche is suspicious of Christian ideals like love and pity. Some types of love are expressions of envy of and/or resentment at the higher qualities of the beloved person and are efforts to appropriate these qualities magically. Nietzsche regards both of these as corrupt motivations for love, and he believes that such symbolic efforts to appropriate human excellence are impoverished substitutes for real self-perfection. Thus, love can express something dark and dubious, and its effect can be to inhibit real self-development. Love can also lead to dishonesty and unrealistic demands if one too readily believes one's idealized vision of the lover. Further, love can lead to insensitivity because the power of the emotion can block one's insight into the deeper psychic dynamics of the beloved or because the lover may refuse to look too deeply. Nietzsche sketches a tougher ideal of friendship as an alternative to personal love (Nietzsche 1883–5/1954: I §14). One's friend may function as an inspiration but not at the cost of one's real self-development. And a true friend is unafraid to challenge one's self-satisfaction and self-deceptions when one's self-development is at stake. The friend is truthful when the lover is silent; the friend challenges the partner to new achievements when the lover celebrates past glories.

Nietzsche's constructive ethic consists of several elements: (1) a general stance; (2) a number of supporting strategies by which that stance is made plausible; and (3) some new values that concretize this stance.

The general stance Nietzsche defends is the enhancement of human capacities – elevating various possibilities of human development to such new heights that contemporary humanity will seem crude and bestial by comparison. This "new" humanity will develop by sublimating the passions of present humanity; indeed Nietzsche's whole enterprise is to harness and organize the actual powers of human beings rather than attempt to define a new "ideal" humanity (Nietzsche 1883–5/1954: I §5). Thus the most critical task for human types and for individuals is to discover the conditions best suited to their development. This may require restricting oneself to pursuing a single capacity for extended periods until it has become second nature and then gradually introducing additional goals. Such personal development requires both discipline and self-knowledge; it requires a great love for one's own future possibilities and a refusal to allow them to lie

fallow. Sometimes one must create ways of making disparate or conflicting passions reinforce, rather than debilitate, one another. Often it involves giving shape and style to one's various capacities; this is Nietzsche's understanding of giving a law to oneself (Nietzsche 1882/1974: §290). But this differs from finding a basic principle; it is more like finding an overall aim that can integrate many of one's talents and aspirations. Through this process of shaping, cultivating, and pruning one becomes a distinctive personality, rather than an anonymous, faceless person.

Three strategies support this general stance: a naturalistic point of departure, an intuitive appeal to health over decadence, and a call to creativity. Nietzsche's position is naturalistic in the sense that he takes himself less to be defining and recommending a transcendent ideal than to be developing and enhancing the capacities and passions humanity already has. He calls this "remaining true to the earth" (Nietzsche 1883–5/1954: I §22). He thinks the capacity to see life as it is and celebrate its real possibilities provides an evaluative standard. The error of religion (and much of classical philosophy) has been to create unreal fantasy worlds that discouraged this-worldly human development. This error is complicated by that of traditional morality, which rebelled against life and against the conditions of human flourishing. Nietzsche's naturalism seeks to reverse these disastrous effects and establish a more promising approach to ethical self-development.

The second strategy is an intuitive preference for health over decadence. Health is a basic condition for the flourishing of all human talents and capacities (Nietzsche 1882/1974: §382). Nietzsche examines its physiological, emotional, and psychic dimensions in both individuals and cultures. A third strategy is to endorse the value of creativity. People are to use their lives to conduct experiments from which future humanity might learn. They should organize the complexity of instincts, drives, and capacities creatively and test particular patterns of such organization. The sheer process of creation has value for Nietzsche. In addition, the most advanced persons bear an additional responsibility: to create new values that can guide the future of human development, that define what "human enhancement" will mean (Nietzsche 1886/1968: §211).

In addition to health and creativity, other fundamental Nietzschean values are life-affirmation, moral strength (a notion that captures what he means by "power"), and deftness. For Nietzsche these characteristics are boundary conditions for living well, in the sense that if one does not embody them one cannot attain one's central defining aspiration or virtue. Also, one would fall outside the wide variety of forms of life that might be promising experiments.

Nietzsche also lauds a variety of particular virtues as plausible beginnings in the task of forging a new humanity. Consider the following list: magnanimity, good-naturedness, gift-giving virtue, justice; then also honesty with oneself, courage, insight, joy, pride, gratitude, reverence, composure, wisdom; and finally, hardness, the willingness to fight for what matters, distance from others, and love of the great possibilities in the future. The first list articulates a way of being with other people purged of all negativity and bad feeling; embodying such virtues would

allow one to contribute to other people without abandoning one's life to them. The second list elaborates Nietzsche's notion of "cleanliness" in relation to oneself, an avoidance of negative passions. It solicits one's finest effort at self-perfection while avoiding gloominess and self-inflation. The final group affirms the importance of conflict in producing self-perfection and extols commitment to the future. These specific values concretize Nietzsche's central ethical stance.

Max Scheler (1874–1928)

Scheler's ethics is rooted in a phenomenological study of the emotions and a critical response to previous ethical theories as well as Nietzsche's attacks on Christianity. He seeks to clarify the import of ethical emotions like love and hate, sympathy and resentment, suffering and shame. He thinks such emotions are the conduits through which humans grasp values, much like the senses are the conduits through which we grasp physical objects and thought is the conduit through we grasp concepts and logical truths. Emotions have their own distinctive order, relationships, and objects; their logic cannot be reduced to the logic of perception or of thought (Scheler 1973: 117–18). Emotions target a specific type of object, values; loving and preferring express an attraction to value. Values motivate all striving or desire. Moreover, Scheler thinks values can be ranked in a hierarchy: sensory, vital, cultural, and spiritual (Scheler 1913–16/1973: 104–10). Humanity participates in all four of these dimensions, and different people experience the values within each level to different degrees of urgency. The distinctive subset of values to which a person is most powerfully drawn defines his or her basic moral tenor (Scheler 1973: 99–111). Scheler's ethics centers on persons, and he regards the realization of a person's basic moral tenor to be among the highest goods.

A close study of the relation of values to human emotions will indicate the flaws in classical ethical theories. Eudaimonism is the view that happiness represents both a natural motive and the highest human flourishing. For Scheler there are four qualitatively different types of happiness (corresponding to the four dimensions of value), but happiness does not define value and nor is it the typical motivation for action (Scheler 1913–16/1973: 328–44). Many eudaimonistic theories identify happiness with pleasure – the lowest form of happiness – because this is the only form of happiness that can be directly sought. Other types of happiness emerge indirectly through realizing certain values. Pursuing sensory pleasures typically compensates for deeper levels of unhappiness (Scheler 1913–16/1973: 345–8). Moreover, such pleasures are not always valuable: addictive pleasures exhibit disvalue as do pleasures taken at another's suffering. Scheler grants that bliss, the highest happiness, is a precondition for grasping the best value possibilities of situations, but insists that it cannot be the direct object of an intention. Actions target values themselves; the better one realizes one's basic moral tenor, the more fully one will be attuned to the higher value possibilities of situations.

Scheler also rejects consequentialism, an ethics of goods, and an ethics of duty. Consequences, like goods, are good only because they embody values. Values explain what existing things and future events are good. Goods often develop, but such developments do not determine value; values determine which developments are good. Goods and purposes, like consequences, are bearers of value, not definers thereof (Scheler 1913–16/1973: 12–23). In general, Scheler thinks that the moral quality of an agent cannot depend on success in realizing consequences because such success is not entirely in the agent's control. Scheler agrees with Kant that moral tenor defines the moral value of a person, but he disagrees with Kant in thinking that moral tenors are quite diverse, and cannot be reduced to a pure desire to do one's duty. Genuine duties also are logically dependent on values, and typically duties are negative – requiring the nonpursuit of evil, rather than the positive achievement of good (Scheler 1913–16/1973: 232–8). Kant's theory suggests that the moral standpoint is impersonal; Scheler disagrees with this, insisting that the realization of one's distinctive personal moral tenor determines moral achievement. Finally, Scheler thinks that reliance on imperatives is necessary only when one is value-blind. With adequate value insight, the bludgeon of duty is unnecessary to motivate action, and utilizing duty to motivate an action is a violation of the person when insight would be sufficient.

Values are distinct from the goods in which they are embodied; they are given differently. A value can be given clearly (justice) while the state of affairs (a just society) that embodies it is given indistinctly, and a good can be given clearly (a work of art) while the value(s) it embodies may be given indistinctly. Values are not destroyed when their bearers are crushed. Thus, values are components of goods, but are not reducible to goods (Scheler 1913–16/1973: 12–23). Positive values are given as to be realized, but this "pull" is not always sufficient to motivate action. Persons are good to the extent that they respond to and realize the highest value possibilities in situations (Scheler 1913–16/1973: 38–44). The value hierarchy is being continuously clarified both by new personal moral tenors and by tenors that exist in diverse cultures and in different eras. All willing and preferring target value realization, not the associated pleasures.

Scheler thinks his ranking of value dimensions (sensory, vital, cultural, spiritual) is intuitively evident, but he suggests properties that support it. Higher values are more enduring, less divisible, provide deeper fulfillment, are the foundation for lower ones, and are more absolute in the sense that transgressing them results in greater guilt (Scheler 1913–16/1973: 90–100). To each level of the four-tiered hierarchy there corresponds a characteristic model person (*bon vivant*, hero, genius, saint), and form of social organization (mass, life community, cultural community, and spiritual community) (Scheler 1913–16/1973: 109, 585). The sector of the value hierarchy to which one is most attuned is one's basic moral tenor. This tenor organizes one's moral life and determines one's goals. Stepping back from one's moral tenor and evaluating it is difficult because it structures one's perceptions (what kinds of objects one sees, and which dimensions

of those objects stand out) and thoughts (Scheler 1913–16/1973: 126–42). It can, however, be altered by conversion or expanded through the influence of personal models.

One of Scheler's major contributions is this emphasis on the importance of personal models in ethical development (Scheler 1913–16/1973: 572–83). Personal models inspire others to realize their own distinctive moral tenors in addition to expanding their value sensitivity. Usually the influence of models is indirect and unintended; this differentiates them from ordinary leaders. Both persons and cultures can learn from each other's moral tenors, and Scheler thinks personal models are gradually clarifying the value hierarchy in its entirety as well as motivating people to realize their best ethical potential.

Scheler supports his theory of value by clarifying the logic of particular emotions (e.g., love, resentment, and shame). Love, for example, is the state in which the highest possibilities of value of the beloved object emerge; it can be directed toward others or oneself (Scheler 1913/1970: 152–61). Self-love is vital to grasping higher levels of value. Personal love of others grasps their basic moral tenor and makes it visible, without intervening to supervise its realization. Resentment results from a sense of impotence characteristic of social groups that lack power and mobility. It inhibits the apprehension of higher values and poisons one's sense of one's own value (Scheler 1912/1961: 43–78). Scheler rejects Nietzsche's claim that Christianity is a resentment-based religion, arguing that Christian love flows from the rich sense of self-value and offers itself in the overflowing radiance characteristic of Nietzsche's higher persons. Scheler agrees with Nietzsche that many specifically modern value stances – for example, humanitarian love, utilitarianism, and relativism – are rooted in resentment. Shame is a natural emotional function that emerges when people suddenly experience a lower dimension of their existence (Scheler 1987: 10–18). It is protective and delays the need to respond to the situation. Shame emerges, for example, when one suddenly becomes aware of one's body when one's spirit or intellect had been the center of attention or when one's particular nature claims attention but only one's stereotypical nature is apprehended. Shame protects the self-value that is being denied or missed, though it does not assert it. It encourages the person to withhold response until the purpose of the denial is grasped (Scheler 1987: 27–36). The experience of value is central to each of these emotions. These sketches illustrate why Scheler thinks emotions provide access to values.

Late in his life Scheler also suggested that the most promising ideal to guide human development in the foreseeable future is that of the complete person or whole person, in which all dimensions of human existence are balanced. To achieve this the current era needs to compensate for the overemphasis on reason, will, and asceticism characteristic of modernity. The Apollinian impulse to order must be compensated with the Dionysian impulse to ecstasy, masculine values with feminine ones, and civilized value perspectives with primitive ones (Scheler 1958: 101–15). This view represents a real change for Scheler; no longer does he stress the highest values or even each person's distinctive moral tenor. Instead, he seeks

a broad incorporation of value by each person, an intermixing of diverse value perspectives that will sensitize people to the fullest spectrum of values.

Jean-Paul Sartre (1905–1980)

Sartre's writings exhibited ethical overtones throughout his career, but he never published a separate treatise on ethics. Indeed, he thought conventional bourgeois morality was bankrupt. Yet he attempted to produce an ethics at many different stages of his career, first in several notebooks and essays written just after the publication of *Being and Nothingness*, then in his biography of Genet (Sartre 1952/1963), and finally in several unpublished manuscripts written after completing the manuscript for the second volume of his *Critique of Dialectical Reason*. The guiding value in all these works is freedom. Its meaning and presuppositions change as his position evolves, but throughout his career Sartre insisted that freedom is a foundation for all human aims and that persons bear responsibility for their actions and their lives. He grounds the value of freedom in the ontological fact of freedom.

In the early period of *Being and Nothingness* the source of freedom is consciousness itself, which perpetually transcends its situation even as it defines itself within such situations. Every situation allows for several courses of action, and making the choice engenders responsibility for the values the choice embodies. Since nothing determines one's response to a situation, one always chooses that response and bears responsibility for those choices (Sartre 1943/1956: 553–6). For Sartre the values one pursues through such choices have no external or rational support. Thus, freedom and responsibility are burdensome, and typically people avoid them through self-deceptive ruses; for example, presuming that social roles determine obligations, pretending that certain values have objective guarantees, or believing that past actions foreclose present choices (Sartre 1943/1956: 55–67). To refuse these self-deceptive ruses, to bear one's responsibility, and to truly author one's long-term projects is to live authentically. At this stage Sartre's notion of freedom is formal in the sense that it seems neither to entail nor to exclude any particular content.

Sartrean freedom in this early phase is also asocial because the primary effect of other people is to produce a dimension of consciousness over which one has no control, the social self. The Other's look petrifies and summarizes one – fashioning an essential nature, molding a permanent sculpture out of a single amorphous moment (Sartre 1943/1956: 259–72). Apart from a brief footnote in *Being and Nothingness*, there is no exit from the struggle among objectifying gazes; either one dominates others by objectifying them, or one is dominated by them by being objectified (Sartre 1943/1956: 268–70, 276–8). Even if one attempts continuous domination, one cannot escape the fact of the social self; it always hovers on the horizon ready for reactivation. The inescapable reality of the social self creates a deep self-division which cannot ever be completely healed.

The overall aim of consciousness, on this early view, is to achieve a synthesis of two types of existence – to be already given passivity (matter) and to be self-creating activity (consciousness) – that cannot be synthesized: in effect to become God (Sartre 1943/1956: 85–95). But since such a synthesis is impossible, Sartre declares that human beings are useless passions. Sartrean conversion requires that people abandon this self-deceptive and fruitless project.

In the middle period, when "Existentialism is a Humanism" and *What is Literature?* were published and his *Notebooks for an Ethics* was composed, Sartre extends his basic position on freedom – showing how the invidious conflict among people can be overcome through reciprocity and how authenticity can function as an alternative to the project of becoming God.

He develops the ontological fact of freedom in three ways. First, he accentuates the weight of one's choices by suggesting that one chooses for everyone; in effect, one's actions function as examples for all to follow (Sartre 1946/1956: 291–2). He thus gives ethical significance to the fact that one's actions have an objective side that is seen and judged by others (Sartre 1946/1956: 293). Second, he suggests that humanity as a whole does not have an essence prior to human action; humanity is something to be made in action both individually and collectively. Historical actions produce the conditions of the lives of present and future generations, and each of one's individual actions reshapes and expresses the fundamental project that informs all one's future actions. In effect, every action contributes to defining and creating human reality (Sartre 1946/1956: 295–8). Finally, since freedom is the foundation of all action, Sartre claims that freedom must be taken as a basic value guiding one's projects and that one's own freedom requires everyone's freedom. Any act which attempts to deny the value of freedom in effect is denying its own conditions, and thus is self-undermining (Sartre 1946/1956: 307–9).

In this period Sartre also grants the possibility of authentic reciprocity, in which people acknowledge and respect each other's freedom. Eventually he will try to show that none can be fully free until all are. One way in which reciprocity can emerge is through common action, in which each person freely adopts the goals of the others (Sartre 1960/1976: 351–63). Another way is that one realizes the dependency of the value and ultimate success of one's actions on others because they must choose to maintain a commitment to one's values when one is dead or infirm (Sartre 1948/1949: 23–42). In addition, Sartre suggests that only through others can one's own pure facticity – one's body and vulnerability – be protected (Sartre 1983/1992: 272–94). This is Sartre's version of Hegelian pure recognition. Each protects what is essentially at the mercy of Others. By choosing to protect each other, the alienated social dimension of oneself is recovered and woven into the fabric of freedom, even if the reciprocating other is the weaver. Still another kind of reciprocity is described in Sartre's study of Genet. Often, one group of people defines another (e.g., Jews, Blacks, Gays) as alien (or Evil) by excluding them from humanity, thus solidifying an experience of themselves as all Good (or truly human). Such attempts abdicate freedom by claiming irrevocable

Goodness. But the open future threatens any such claim. To the extent that one abides with one's freedom, one must acknowledge the ambiguity that qualifies both oneself and others. This allows one to accept a fundamental identity with others and reduces the urge to exclude and stigmatize them.

Sartre further explores the nature of authenticity by clarifying the process of achieving it, which he calls "conversion." Conversion involves learning to fully acknowledge one's freedom and responsibility, instead of escaping them through self-deceptive maneuvers. Conversion is an active correlate of purifying reflection. In purifying reflection one discovers the major Sartrean truths: that human existence is contingent and without support, that values have no external or objective guarantees, that life is a series of freely chosen projects, and that each situation opens new possibilities. This emerges in a flash of recognition. Conversion simply abides with this experience, continuously living in accord with it. Instead of trying to imagine oneself outside history, one lives historically; instead of living oblivious to one's body, one accepts one's embodiment (Sartre 1983/1992: 471–514). But one also refuses to be engulfed by the given aspects of life; one refuses to be reduced to one's body, and one responds to the challenges of history. In effect, one lives inauthentically if one allows oneself to become reduced to facticity or if one tries to escape it entirely. One thus commits oneself to specific goals and values, but periodically reviews those commitments. By assuming one's freedom, one lives both within and ahead of history, both within and ahead of one's body, one's past, one's situation, one's social definition, and one's death.

In the later period of *The Critique of Dialectical Reason*, the locus of freedom shifts from consciousness to praxis. Praxis still transcends given situations, but now it is internally shaped by those situations and mediated by a variety of social and material conditions. Thus, Sartre better incorporates the historical weight of situations – the manner in which they make their own demands and bear their own inertia. This inertia derives from the conflicting projects of existing social groups, from the resources and technology handed down by past generations, and from scarcity. In this period Sartre acknowledges that history makes individuals as much as individuals make history. Further, he recognizes that profound social changes are required for most people to be free in the mundane sense of not being dominated by need. In addition, Sartre defines a social process of achieving reciprocal recognition through group action – often revolutionary action. The group-infusion emerges when members of a "street action" find themselves aiming at the same goal (often, defending themselves against an attack) and thus searching for an effective means. Here group praxis and individual praxis interpenetrate one another: each member is end and means for the others; each recognizes the other's action as her own; and each directs others only insofar as others also give direction (Sartre 1960/1976: 374–83). Sartre also claims that group action typically produces social structures that undermine the spontaneous freedom that a group initially discovers by pursuing collective action. The question thus becomes whether this kind of group reciprocity can be sustained over time. Even if it can, groups never escape an ominous background, for example, an implicit fear that enforces

every member's pledge and sustains their brotherhood (Sartre 1960/1976: 428–44).

Sartre also discovers some other dimensions of alienation in this period. Scarcity produces alienation in that it turns humans against one another in competition. In addition, the current technology embodies the purposes of its creators, and thus current users must conform themselves to these purposes as they use it. To use a car is to live in an entire practical world created by the car, and this shapes the user's own goals and aspirations (Sartre 1960/1976: 161–96). Finally, there are the processes by which the group-in-fusion becomes an institution and thus common praxis returns to a condition of social seriality – the condition in which each is other to all the others, and in which only numerical relations exist between them (each is but one of many in a series) (Sartre 1960/1976: 664–70). A full overcoming of human alienation would require transforming these conditions. Sartre claims that these conditions make people subhuman. His dialectical ethics clarifies how people can combat the subhumanity in themselves and others.

Sartre's ultimate social aim is clear: a world in which everyone contributes to making history through common group praxis and where each recognizes all the others as history is created. Here each person is free in and through everyone else's freedom (Sartre 1983/1992: 468–71). To reach this goal, persons must give their present praxis this historical goal, finding ways to transform the subhumanity of themselves and others. The very condition of praxis foreshadows this aim, just as authenticity is foreshadowed in the dynamic of consciousness. Praxis can flee this implicit goal, however, by maintaining the present inhuman conditions; it does this by adopting and maintaining the values of the dominant order. To transform ourselves from subhuman to truly human, we must alter the historical conditions that create that subhumanity, including scarcity and ossified institutions. Sartre thinks the impetus for this transformation must come from the exploited and suppressed groups and classes, and the potential for such transformation is ever present, but achieving it is never guaranteed. "Ethics" now becomes the effort through which this potential freedom of all is pursued historically. Whether such efforts succeed, are defeated, deviate from their central goal, or undermine themselves depends on human choices and historical conditions. *The Critique* explores the many ways by which historical action can fail. The unpublished dialectical ethics suggests that failure is not inevitable.

Emmanuel Levinas (1906–95)

Levinas offers a new type of ethical theory, one which roots ethics in a fundamental relationship to other people. For Levinas this relationship establishes the fundamental ethical orientation, but the other person remains transcendent, alien, and inassimilable. Buber's I–Thou relationship offers a similar type of theory, but his ethically constitutive interpersonal relation is quite different. Levinas insists on an *asymmetry* between self and other while Buber's I–Thou relation creates symmetry

(both become "Thous" for each other). Levinas insists on the other's absolute transcendence, which prevents symmetry and reciprocity (Levinas 1961/1969: 39–40, 51–2). Levinas's other becomes manifest through the human face which both commands one (not to harm) and solicits one's aid; to respond to the other's face is to be responsible to the other and for the other (Levinas 1961/1969: 197–200).

Levinas challenges many Greek assumptions in Western thought, including the primacy of Being and knowing. In Western thought "being" is typically conceived as "what is present" – as the given. And knowing is a matter of assimilating or integrating or possessing Being. One challenge to this analysis of Being was issued by Heidegger, who drew attention to our practical relationship with tools and thus revealed a different kind of knowing (knowing how) and being (tools vs objects). But Levinas suggests that tools are simply given in a different way and claims that Heidegger's concept of Being never really escapes the orbit of the notion of presence. For Levinas the other's *proximity* is prior to "presence" and makes it possible. The face that makes ethical demands is neither an object that comes to presence nor a tool. The face is a mystery that defies assimilation, and is thus beyond being and knowing in the ordinary sense. Also, in knowledge a sovereign subject incorporates what is known, but the Other's command and helplessness shatter one's sovereignty and challenge one's efforts to possess (Levinas 1947/1987: 87–90). For these reasons the other is beyond knowledge. Levinas's ethics explores the implications of this mysterious other, which challenges the primacy of both ontology and epistemology. Levinas insists that ethics is first philosophy: it reveals the impossibility of possession, pure presence, and assimilation.

The task then becomes clarifying this fundamental relationship to others. No synthesis with the other is possible; thus no totality (a whole that integrates its parts) can unite self and other. Most interpersonal ideals derive from this notion of totality. For example, Hegelian reciprocal recognition produces a symmetrical harmony that binds the participants together into a new whole. This Hegelian model of unity is completely rejected by Levinas. Instead of inviting such harmony, the Other shatters one's self-possession and self-enclosure, rendering one vulnerable and exposed. To face the other is to answer an already existing summons, demand, and obligation. Others interrogate one; they are the source of one's sense of duty (Levinas 1961/1969: 82–4). For Heidegger the fundamental feature of human being is to be at issue; for Levinas the other (not Being or *Dasein's* essence) is the source of this decentering experience. Responding to this challenge is one's first ethical act. One's relationship to the other is like a relationship to infinity, which is perpetually beyond experience (Levinas 1961/1969: 48–52). Like time, the other is a dimension in which one is immersed but which one cannot circumscribe, delimit, or define. Like the genuine future, which is forever surprising, the other's face explodes one's own plans and purposes (Levinas 1947/1987: 79–81).

Levinas struggles to create appropriate metaphors to clarify this ethical relation to the other and explicate its importance, often using different metaphors in

different books. In *Time and the Other* he uses physical pain, death, paternity, and heterosexual desire as useful analogies for this relation. Excruciating pain renders one impotent and shatters one's expectations – emphasizing one's helplessness – but one's response is immediate and pointed (Levinas 1947/1987: 68–71). This experience captures the urgency and intrusiveness of the other's call. Death is an ever-present possibility that cannot be assimilated or known, cannot be mastered or escaped. Moreover, it is not similarly threatened: one cannot rebel and dominate death. Facing the ethically fundamental other is like facing death – equally unfathomable, inassimilable, and relentless (Levinas 1947/1987: 71–7). In heterosexual erotic love, the other remains a mystery, distinct from oneself yet familiar; by caressing the beloved one yields to this mystery without trying to assimilate it. To caress is not to know or master another, but to meet him or her in a way entirely different from knowledge. In the caress one loses and risks oneself, and this risk is an essential element in ethical response (Levinas 1947/1987: 87–90). Finally, in paternity one is related to someone wholly other, but who carries one's legacy. The son offers the father a way to survive death. Fathers are reflected in sons, but sons are wholly distinct from and transcend their fathers (Levinas 1947/1987: 87–90). These metaphors help Levinas articulate the relationship to others that grounds ethical responsiveness.

In *Totality and Infinity*, Levinas challenges the concept of totality, replacing it with a relation to infinity. Just as self and other cannot be integrated, politics and ethics cannot be harmonized. Ethics concerns only this fundamental relation to the other; when more than one person is involved, each becomes a third with claims and interests to be weighed. Such weighing of competing interests produces the qualitatively different realm of politics and history. In politics everyone is treated as an object – each equal to the others, but in ethics self and other are never equal. In politics one's obligations to others can be limited, but not in ethics in which the other's command is absolute. One's ethical relation to others is not altered or transformed by history; it is not relative to history; indeed, it can be used to judge historical agents and eras (Levinas 1961/1969: 21–6). In this book Levinas develops his central conception of the face of the other. The face is exposed, naked, and defenseless, but it is also upright and commanding. Both higher and lower than oneself; both master and supplicant, the other's primary injunction is not to harm.

In *Otherwise Than Being or Beyond Essence*, these metaphors are reworked, and new ones emerge. Central is a notion of responsibility that involves substituting oneself for the other; Levinas now suggests that one bears the burden not only of one's response to the other, but also of the other's own actions. Thus Levinas moves beyond responsibility as responsiveness to bearing the other's responsibility. One answers for all the others (Levinas 1974/1981: 113–18). Such responsibility defines the structure of ethical subjectivity. Levinas realizes this assertion seems excessive, beyond what can be reasonably expected. But ethics is not a matter of reasonable expectations. It develops the implications of one's fundamental

relationship with infinity. To assume this excessive responsibility is to leap beyond life and death, to transcend a life that seeks only to sustain itself.

For Levinas, there are no founding values and no basic principles; the core relationship to the other is the basis and source of all ethical obligations. Moreover, there are few arguments. Many of his formulations echo religious metaphors, but he recasts and revitalizes them. He paints conceptual pictures. He tries to show how his apparently abstract notions intersect with a wide diversity of experiences. Other interpretations of such experiences might be possible. But like those of many phenomenologists, his claims embody experiential insights. He is a clever critic of both Hegel and Heidegger, and is clearly the spiritual father of philosophies of dispersion, both in French feminism and in poststructuralism. He attempts to articulate the inexpressible.

Some Implications

In the introduction I indicated some features of Continental ethics that emerge from a contrast with contemporary analytic ethical theory. I will return to this theme circuitously, first by exploring some apparent contrasts among the Continental ethical theories briefly sketched in the previous section, then by indicating some deeper similarities that will develop my earlier remarks and perhaps raise some challenges to the approaches to ethical theory described elsewhere in this book.

Three disagreements among the Continental thinkers seem apparent: (1) some believe that values are in some sense given (Hegel, Scheler, and Levinas) while others believe that they must be created (Nietzsche, Sartre); (2) some seek reciprocity and mutual recognition as their primary social ideal (Hegel, Scheler, Sartre) while others defend the value of a more agonal, asymmetrical relation to others (Nietzsche, Levinas); (3) some take freedom to be the fundamental and most defensible value (Hegel, Sartre) while others question the value of freedom and instead seek to defend a richer set of substantive ideals (Nietzsche, Scheler, Levinas). Each of these apparent oppositions deserves further examination.

The first contrast is less stable than it first appears. The kinds of givenness articulated by these thinkers are quite diverse. Hegel believes that ethics begins with the values of the existing culture; they provide the basic substance of ethical life. Scheler thinks that the emotions provide a window to an independent sphere of values. Levinas thinks the ethically fundamental other transcends, challenges, and calls each person to account. But each of these thinkers believes that the ethical agent must transform or respond to these given elements. Hegel thinks the existing values and institutions of the culture must be rationally interrogated and harmonized, pruned and refined, and above all tested for their contribution to mutual recognition. Scheler thinks that insight into the value hierarchy can always be broadened by moral visionaries (and exposure to alien cultures) and that

everyone's creative task is to discover their moral tenors (which are often concealed by others' expectations or their own preoccupations) and to richly embody their distinctive values. Finally, Levinas thinks the ethical task is finding creative responses to the insistence of the other. The other's call is only a point of departure for serious ethical life; one's response is its core. So each theorist who takes fundamental values to be given also believes that living ethically requires a creative response to the givens, not merely an assimilation of them.

In contrast, the theorists who argue for the importance of value creation also believe this process must be guided by the givens of human existence and historical possibility. Sartre's concept of conversion involves learning to acknowledge the ambiguous reality of the human condition. Then one must commit oneself to values that will take the whole of humanity in promising directions. Sartre's ideal requires finding concrete ways to realize everyone's freedom amidst the crosscurrents and inertia of the historical situation. In addition, Nietzsche's hopes for creating perfected human types depend on lucid knowledge of human psychology and cultural dynamics and seek to find a role for existing virtues that stand the test of his rigorous revaluations. Both thinkers thus defend meta-values or boundary conditions within which new value creation is to occur; they thus indicate promising general directions.

In fact all the theorists acknowledge some dimension of value givenness and some demand for value revision, transformation, or creation. Thus, this initially sharp opposition in fact becomes a difference of degree and emphasis. This suggests the realism/constructivism debate is less clear and coherent than it initially appears. Many have felt that if only values can be given objective guarantees, the work of ethical theory will then be complete. The lesson here is that various kinds of objectivity play significant roles in ethical theory, and that serious ethical thought requires equally penetrating reflection on the individual's uptake of the objective elements. Most Continental ethicists accept both an objective element in and an essential individual contribution to genuine ethical life.

The second opposition concerns the ideal that informs social life: pure recognition vs asymmetrical challenge. Though both Hegel and Sartre take reciprocal recognition as an ideal, both assert that conflict and struggle are necessary moments on the path to achieving it and that such struggle can make positive contributions to its emergence. Both master and slave learn essential lessons in Hegel's *Phenomenology*, and historical struggles provide the vehicle for the formation of groups that eventually achieve mutual recognition in Sartre's *Critique*. On the other hand, theorists like Nietzsche and Levinas – who extol the value of conflict or challenge in interpersonal life – do not sanction domination or oppression. The conflict produces a mutual enhancement or awakening that does not entail equality (for Nietzsche) or assimilation (for Levinas). Both of these thinkers value the way in which persons unsettle and challenge each other, thereby overcoming ethical stagnation and lassitude. Thus, here too the opposition between the two sides is a matter of degree and nuance. Both groups find an ethically productive role for conflict (negativity), and both acknowledge the virtue of some kind of mutuality

or reciprocity as well. Even the seemingly most radically opposed thinkers, Hegel and Levinas, may be miscast. Hegelian recognition does not entail the assimilation of others and transcendence of their distinctiveness or their power to challenge. And despite his insistence on the commanding quality of the other, Levinas embraces a pacific ideal. These reflections show both the importance of intersubjectivity in Continental ethics – especially insofar as personal relations can be a source of ethical transformation – and the radically different ways in which one's ethical relationships with others are conceived (the egoism/altruism opposition is superseded by this conflict–harmony mix; see later).

The third dimension opposes thinkers valuing freedom to those seeking the realization of more substantive ideals. Once their positions are fully explicit, however, the concepts of freedom governing the ethics of both Sartre and Hegel are multifaceted, and – more importantly – rarely stand alone. Hegel expects free persons to breathe life into their culture's roles and values. Sartre expects people to commit themselves to their own chosen ideals, so long as they remain consistent with the development of social freedom. Also, they both argue that a variety of additional conditions are necessary to achieving freedom. Thus, on closer analysis, they are committed to realizing a broad range of conditions and facets of freedom. On the other side, though Nietzsche is sharply critical of many conceptions of freedom, he values personal coherence and strength and frequently calls them "freedom." Also, while Scheler thinks persons are circumscribed by their basic moral tenors, these tenors nonetheless can be expanded through exposure to distinctive personal models, and he would acknowledge that living in accord with one's basic moral tenor is what many people mean by "freedom." Even Levinas – who generally stresses the demands others impose – accepts the value of a sphere of individual freedom and domesticity that allows people to respond to the other's call effectively. The real issue here concerns less the positive value of freedom or the range of substantive ideals embraced, but the meta-value or boundary values that govern each thinker's theory (Hegelian recognition, Nietzschean vitality, Schelerian spirit, Sartre's community of ends, Levinas's responsibility to and for other). A kind of freedom finds its place in each theory, as do a variety of substantive ideals. But the meta-values determine the distinctive aspiration and direction of each theory; they indicate the real substantive differences among the theorists.

Examining each of these apparent oppositions has provided some clarification of the real issues at stake in Continental ethics. I shall conclude with a discussion of five features that establish the distinctive conceptual space of Continental ethics. In the process, I will indicate some ways in which analytic ethics may be misguided. To fully establish the superiority of the Continental "space" of ethical thought would require a separate essay, but at least its outlines can now be sketched.

First, one central assumption motivating ethical theory in the analytic tradition is that the function of ethics is to combat the inherent egoism or selfishness of individuals. Indeed, many thinkers define the basic goal of morality as

"selflessness" or "altruism." Almost immediately this assumption gives ethical theory a policing function, which is one of the reasons many Continental thinkers are skeptical of everyday "morality." For a multitude of reasons – partly because they respect the power of socialization, partly because they believe in a more complex human psychology, partly because they reject the claim that selfishness is always invidious – these thinkers bypass this obsession with selfishness. This does not mean they ignore the ethical significance of interpersonal relations and inter-subjectivity; they simply think that the egoism/altruism opposition obscures the function of ethics. For them the goal of ethical theory is awakening people to their higher selves, helping individuals find the inner and outer resources that will enable ethical flourishing. The goal, in effect, is not to police or circumscribe, but to awaken and enliven – to energize people to ethical creativity.

Second, and correlated with this point, Continental ethicists eschew theories of obligation and duty. Scheler expressed one central reason for this. Duties are primarily negative and typically dependent on values. Continental thinkers explore the underlying values and articulate the positive motivation for pursuing them. Their task is not to defend a set of regulations that everyone "must" obey – as if people were little children and ethical theorists were schoolmarms. Rather they seek to produce an awakening – as if people all-too-often sleepwalk through their lives. Duties, imperatives, rules, and regulations only deepen this daze (which one might call ethical death), and for that reason they are suspect. This is another reason why I suggested at the outset that Continental thinkers think "morality," narrowly construed, is itself immoral. Even if such moralities were to succeed, at best they would produce befogged herd people (as Nietzsche suggested). One cannot be inspired by interdictions, but one can be moved by higher values. Finding a way to light this "spark" is one central task of Continental ethical theories.

Third, reason has a different role in Continental ethical theory than in analytic theories. Initially, "reason" may be defined more broadly (to include, for example, Hegel's dialectical reason), or it may be construed to require a more psychological penetration and subtlety in grasping specific human types and individuals (such as Nietzsche sought with his "genius of the heart"), or it may be reinterpreted entirely as a "logic of the heart" (such as Scheler claimed to discover). But rarely is practical reason a matter of deductive arguments with major and minor premises; nor is it a matter of weighing competing considerations on some commensurable dimension (e.g., happiness or utility); nor is it a matter of providing an indubitable rational justification for some first principle of justice or benevolence. Continental theorists do rationally defend their own *theories,* but one implication of most of their theories is that cognition is only one element of living ethically, and the kind of cognition involved requires subtle analysis. Some might be tempted to claim that these theorists are noncognitivists, but in fact they are all seeking to articulate a different kind of cognition, one more suited to the real requirements of serious ethical life.

Fourth, Continental ethicists rarely seek or defend a single set of universal values to which everyone is expected to aspire. They do not embrace relativism, and they do not think everyone's "values" are equally good. But they acknowledge that they are different human types, and that different ethos may better realize the highest possibilities of each type. Even Scheler – who was a realist about values – insists that different personalities have distinct basic moral tenors and that these must be respected and enhanced. Hegel might recommend different educational institutions and communal structures for those pursuing very different roles (e.g., political leader, householder, artist, farmer). Nietzsche most assuredly had different guidelines and goals for different human types. They all recognize the value of discovering and defending meta-values or general directions for human aspiration, but at best that is the first step. The serious ethical thinking comes when one examines the manifold ways of pursuing these directions or instantiating these meta-values. Continental theorists above all seek genuine ethical achievement. This is why they examine the many ways in which people get derailed and the conditions that can improve their odds. Contrast this with the central project of analytic ethics – to produce a rational foundation for a small set of universal principles – and the differences between the traditions become evident.

Fifth, Continental ethical theories nearly always define and respond to the larger cultural stakes of their historical eras. Hegel reacts to the French Revolution, as well as to Kant. Nietzsche responds to the rise of mass culture and to Wagnerian art, as well as to Schopenhauer. Scheler responds to the vicissitudes of World War I, as well as to Husserl and Brentano. Sartre and Levinas in different ways respond to the horrors of World War II, as well as to Heidegger. Their ethical theories do not succeed or fail because of these historical conditions, but these larger cultural developments are never far from their minds. So much of analytic ethics seems oblivious to the major issues of the larger culture; it almost prides itself on its diffidence toward such developments. Only some of the revolutionary approaches indicated above have registered some awareness of current cultural realities. None of the Continental theories have become outdated because these cultural contexts have receded. Their analysis of the culture's core issues was sufficiently penetrating that most remain alive today. People are not abstract place-holders outside of all time and place; they are historically situated. Explorations of their self-enhancement must take account of this situatedness if they are to be effective.

Finally, the most distinctive feature of Continental ethical theory is that it is not really isolated from metaphysics, philosophical anthropology, philosophical sociology and psychology, or epistemology. One can discuss the ethical elements of Continental theories separately, but they are not really produced independently. They are woven into a systematic treatment of a range of issues, and the confidence one develops in a particular ethical perspective is often connected to one's confidence in the overall system. Analytic philosophy has taken another path. It treats the separate areas of philosophy as distinct specializations, and the issues

within each specialty often become detached from any larger systematic vision. Perhaps the reason Continental ethics has attracted so little interest among ethicists in the analytic tradition is that they find such larger systematic visions alien. But the deeper reason is that the project, perspective, and assumptions informing Continental ethics really are different. Perhaps by clarifying some of these differences, I have created a way to appreciate the force and challenge of Continental ethics. Perhaps I have also raised some questions about the value of analytic ethics as it is all-too-often pursued.

References

Boyd, R. (1988) "How to Be a Moral Realist," in *Essays on Moral Realism*, ed. G. Sayre-McCord, Ithaca, NY: Cornell University Press, pp. 181–228.

Foot, P. (1978) *Virtues and Vices*, Los Angeles: University of California Press.

Hegel, G.W.F. (1807/1977) *Phenomenology of Spirit*, trans. A.V. Miller, Oxford: Clarendon Press.

Hegel, G.W.F. (1821/1967) *The Philosophy of Right*, trans. T.M. Knox, Oxford: Oxford University Press.

Levinas, E. (1947/1987) *Time and the Other*; trans. Richard Cohen, Pittsburgh: Duquesne University Press.

Levinas, E. (1961/1969) *Totality and Infinity*, trans. Alphonso Lingis, Pittsburgh: Duquesne University Press.

Levinas, E. (1974/1981) *Otherwise Than Being or Beyond Essence*, trans. Alphonso Lingis, The Hague: Martinus Nijhoff.

MacIntyre, A. (1981) *After Virtue*, Notre Dame, IN: University of Notre Dame Press.

McDowell, J (1985) "Values and Secondary Qualities," in *Morality and Objectivity*, ed. T. Honderich, London: Routledge & Kegan Paul, pp. 110–29.

Murdoch, I (1970) *The Sovereignty of the Good*, London: Cambridge University Press.

Nagel, T. (1979) *Mortal Questions*, New York: Cambridge University Press.

Nietzsche, F. (1881/1982) *Daybreak*, trans. R.J. Hollingdale, Cambridge: Cambridge University Press.

Nietzsche, F. (1882/1974) *The Gay Science*, trans. Walter Kaufmann, New York: Random House.

Nietzsche, F. (1883–5/1954) *Thus Spoke Zarathustra*, trans. Walter Kaufmann, New York: Viking.

Nietzsche, F. (1886/1968) *Beyond Good and Evil*, in *Basic Writings of Nietzsche*, trans. Walter Kaufmann, New York: Random House.

Nietzsche, F. (1887/1968) *The Genealogy of Morals*, in *Basic Writings of Nietzsche*, trans. Walter Kaufmann, New York: Random House.

Nietzsche, F. (1888/1954) *Twilight of Idols*, in *The Portable Nietzsche*, trans. Walter Kaufmann, New York: Viking.

Noddings, N. (1986) *Caring: A Feminist Approach to Ethics and Moral Education*, Berkeley: University of California Press.

Nussbaum, M. (1990) *Love's Knowledge*, New York: Oxford University Press.

Sartre, J.-P. (1943/1956) *Being and Nothingness*, trans. Hazel Barnes, New York: Philosophical Library.
Sartre, J.-P. (1946/1956) "Existentialism is a Humanism," in *Existentialism from Dostoevsky to Sartre*, trans. Philip Mairet and ed. Walter Kaufmann, Cleveland, OH: World Publishing.
Sartre, J.-P. (1948/1949) *What is Literature?* trans. Bernard Frechtman, New York: Philosophical Library.
Sartre, J.-P. (1952/1963) *Saint Genet, Actor and Martyr*, trans. Bernard Frechtman, New York: George Braziller.
Sartre, J.-P. (1960/1976) *Critique of Dialectical Reason*, vol. 1, trans. Alan Sheridan-Smith, London: New Left Books.
Sartre, J.-P. (1983/1992) *Notebooks for an Ethics*, trans. David Pellauer, Chicago, IL: University of Chicago Press.
Scheler, M.(1912/1961) *Ressentiment*, trans. William Holdheim and ed. Lewis Coser, New York: Free Press.
Scheler, M. (1913/1970) *The Nature of Sympathy*, trans. Peter Heath, Hamden, CT: Archon Books.
Scheler, M. (1913–16/1973) *Formalism in Ethics and a Non-Formal Ethic of Values: A New Attempt Toward the Foundation of an Ethical Personalism*, trans. Manfred S. Frings and ed. Roger L. Funk, 5th rev., Evanston, IL: Northwestern University Press.
Scheler, M. (1958) *Philosophical Perspectives*, trans. Oscar Haac, Boston: Beacon Press.
Scheler, M (1973) *Selected Philosophical Papers*, trans. David Lachterman, Evanston, IL: Northwestern University Press.
Scheler, M. (1987) *Person and Self Value: Three Essays*, trans. Manfred Frings, Dordrecht: Martinus Nijhoff.
Slote, M. (1992) *From Morality to Virtue*, New York: Oxford University Press.
Stocker, M. (1990) *Plural and Conflicting Values*, New York: Oxford University Press.
Taylor, C. (1989) *Sources of the Self*, Cambridge: Harvard University Press.
Wallace, J. (1978) *Virtues and Vices*, Ithaca, NY: Cornell University Press.

Index

The Blackwell Guide to Ethical Theory, Second Edition. Edited by Hugh LaFollette and Ingmar Persson.

Index compiled by Frank Pert