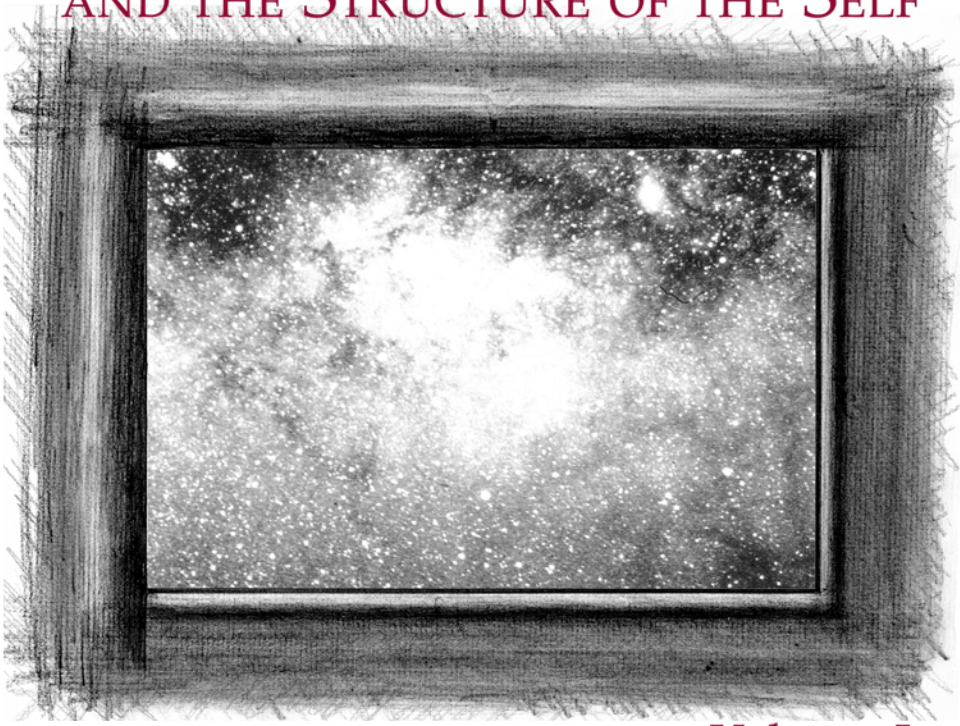
Rationality and the Structure of the Self



Volume I: The Humean Conception

With a new Preface to the Second Edition

Adrian M. S. Piper

RATIONALITY
AND THE
STRUCTURE OF THE SELF

Volume I:
The Humean Conception

Adrian M. S. Piper

2nd Edition

First published 2008
2nd Edition 2013

ISBN #978-3-9813763-2-6

Cover by Adrian Piper

To the Memory of John Rawls

RATIONALITY AND THE STRUCTURE OF THE SELF

Contents of Volume I: The Humean Conception

Summary Contents

RATIONALITY AND THE STRUCTURE OF THE SELF

Contents of Volume I: The Humean Conception

Detailed Contents

List of Figures

Preface to the Second Edition

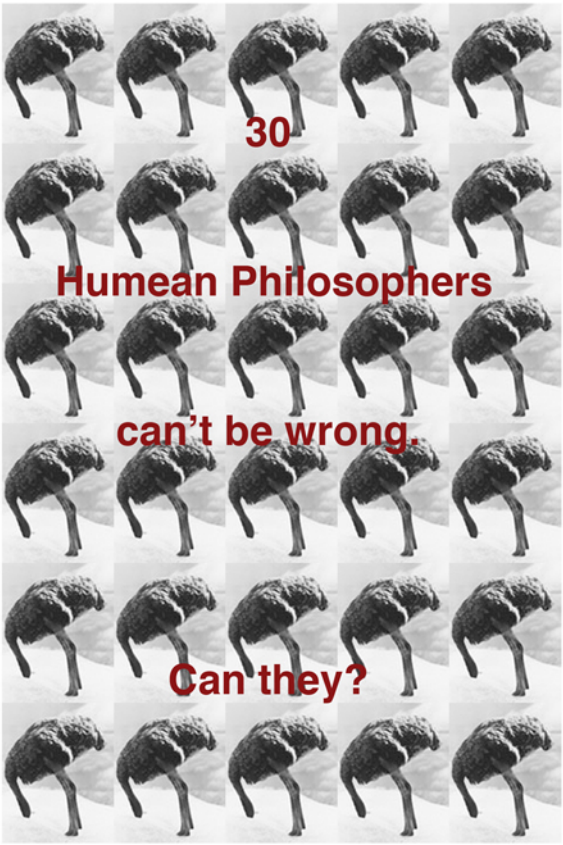
Rationality and the Structure of the Self has always had a curious history; indeed, 34 years' worth to completion. But those were relatively uneventful, compared to its publication history, which has only grown curiouser and curiouser. This fifth publication anniversary, marked by a reformatted and redesigned second edition, is an opportune moment to review and take stock.

When Cynthia Read first solicited *Rationality and the Structure of the Self* for Oxford University Press in the early 1980s, it was a longish, one-volume manuscript that – as I predicted at the time – promised to grow. She apprised me of OUP's traditional sympathy for multi-volume projects (by Frances Myrna Kamm, Bimal Krishna Matilal, Alexander Murray, Werner Jaeger, Wayne Waxman, Terence Irwin and Derek Parfit, to name a few recent examples). So in the late 1990s, I kept my promise to get back in touch when it was close to completion. By then it had grown to four volumes. Peter Momtchiloff insisted that I cut it down to two. I did that. Then he insisted that I cut it down to one. I refused, and withdrew.

Terry Moore of Cambridge University Press solicited *Rationality and the Structure of the Self* in the early 1990s. I brought it to CUP in the early 2000s, and stated at the outset my refusal to cut it any further. I worked with Beatrice Rehl. She was the best editor I could have wished. She understood and respected the interconnection of both volumes, the impossibility of marketing each as a completely independent work, and even my stubborn refusal to further reduce the size of either one.

But Beatrice was even better than that. Because *Volume I: The Humean Conception* is very critical of a conception of the self that virtually everyone, both in philosophy and in the social sciences, takes for granted, it was extremely difficult to find reliable readers for this volume. More than thirty people simply refused to read it, and Beatrice refused to countenance the impertinent poster I designed in order to exploit the marketing potential of this remarkable fact (see Figure 1, next page). A few of my colleagues wrote reader's reports that were so mad-dog, chewing-up-the-rug savage that they subverted their own credibility. For example, one fulminated against its purported failings at very great length, without bothering in any instance to cite the text. Another fabricated objectionable text against which to fulminate, in the apparent certitude that Beatrice had not bothered to familiarize herself with the text I actually wrote. A third, so thinly disguised as not to have needed to bother with the pretense of anonymity, objected to my having neglected to discuss her recent book.

Any other editor would have used such reports as a convenient excuse to get rid of Volume I entirely, and demand that I publish *Volume II: A Kantian*



Find out at adrianpiper.com/philosophy.shtml

We decided to publish Piper's

Rationality and the Structure of the Self: The Humean Conception,

a comprehensive critique of the dominant conception of the self in the humanities & social sciences, at the Adrian Piper Research Archive website because 30 Humean philosophers declined Cambridge University Press' request to read and review it for publication. Read it there yourself and decide whether or not the Humean conception of the self is worth critiquing. And order ***Rationality and the Structure of the Self: A Kantian Conception*** from Cambridge University Press to explore Piper's proposed alternative.

RATIONALITY AND THE STRUCTURE OF THE SELF
is a joint publication of Cambridge University Press and the Adrian Piper Research Archive.

Figure 1. Proposed Promotional Poster (2007)

Conception either separately or not at all. Beatrice could have done that, but she did not. Instead she spent a great deal of time and money finding readers for both volumes whose word, though critical, could be trusted. Both volumes are very much improved for the rigorous, constructive criticism and encouragement her chosen readers finally supplied. My debt to her and to them is very great. It was a privilege to work with an editor of this calibre.

But CUP's review procedure is unusual in requiring yet a further round of vetting: Each volume also had to be independently read and approved by the Cambridge University Press Syndicate, a group of eighteen Cambridge University professors from different disciplines who pass judgment on each manuscript which CUP's editors submit for publication. That both volumes of *Rationality and the Structure of the Self* survived this highly ramified gauntlet of specialized professional evaluation reinforces my belief in its worth.

After both volumes had been fully and formally approved for publication by academic scholars professionally trained to make such judgments, CUP's marketing department then demanded that I cut 100 pages – any 100 pages – from each volume, in order to sell them more easily. Beatrice had agreed in writing not to require this. But it is CUP's marketing department, not its editors or syndicate of scholars, that finally determines what CUP publishes and in what form. Of course the resulting books would not have been the ones that the CUP Syndicate had approved. I refused, withdrew, and published both volumes at my website.

This is what happens when you break a promise to a Kantian.

Although CUP's vetting procedure is unusually demanding, its ultimate deferral to the financial bottom line is not unusual at all. The reality is that the economic climate for all print publishers, but particularly for academic print publishers, has been extremely difficult and getting steadily worse over the last decade. Pig-headed authors such as myself do not help the situation. Some publishers are forthright and transparent about these limitations. Others try to make a virtue of necessity, and to convince their authors that these limitations are, indeed, a virtue. As I accept only those limitations dictated by the imperatives of the work itself, I have sought virtues elsewhere.

I did not write *Rationality and the Structure of the Self* in order to make a profit. But I have derived very great profit indeed from its instant accessibility to anyone beset by even a momentary flicker of curiosity about its contents. Electronic, open-access self-publication has also done much more to bring it to public attention than a traditional print publisher's contract would have allowed. Full-page advertisements in *The Proceedings of the American Philosophical Society*, *The Journal of Philosophy, The Philosophical Review, Mind, Ethics, Political Theory, The European Journal of Philosophy,* and *Economics and Philosophy* have secured its place in the historical record. And advertising it on

the Philosophy in Europe E-List[1] has inadvertently generated some very heated debate about having done so.

Granted: disagreements with the actual arguments of *Rationality and the Structure of the Self* have not been "aired where they should be, in the arena of scholarly debate," as one of its anonymous readers had expected. Indeed to my knowledge, it has not received a single mention, much less a review, in any academic forum, conference, journal or book in the five years since its first publication; and it may well have to wait for many people to die, including me, before it gets one. I can live with that.[2] For in the end, we all die. Then all that is left is the work, and all that matters is its quality.

But in the meantime, this over-my-dead-body collective public disregard has enlivened a thriving private interest in both volumes at my website. Off-the-record magnanimous comments have also provided cardiopulmonary life support. And a proliferation in recent years of talk, conference, journal, and edited collection topics concerning the self, self-deception, desire, reasons, rationality, and the Humean model of motivation has had an equally pleasant resuscitating effect. Perhaps I will rise from the grave. In any case, these developments at least embalm the project in a regenerative admixture of edginess and scholarly significance.

Rationality and the Structure of the Self also has manifested a different kind of significance. In effect, it has been functioning as a litmus test of the theory of professional power dynamics introduced in Chapter I. Formulated in 1998, that theory best explained the data of my experience and observations in the field of academic philosophy:

> It is because rational philosophical dialogue recognizes no professional hierarchy that other, extra-philosophical or even anti-philosophical measures must be invoked to maintain it under circumstances in which hierarchical status is the surest index of professional survival. … In this traditional hierarchy, with few exceptions, … novices, newcomers,

---

[1]See Adrian Piper, "Re.: Self-Advertisements," posted by Philosophy in Europe PHILOS-L@liverpool.ac.uk on Saturday October 4, 2008, at 15:01. Archived at http://listserv.liv.ac.uk/archives/philos-l.html. With over 7,000 subscribers in 57 countries, plus duplication to several additional global redistribution lists, the Philosophy in Europe e-list is the largest philosophy mailing list in the world.

[2]I argue in Volume II that when most people want to do something, they find a reason to do it; whereas when they want not to, they find a reason not to. So the deafening silence has not moved me to seek explanations for it. But some have been pressed upon me nevertheless. Impromptu public remarks about the project include "pretentious," "presumptuous," and the opinion that *Rationality and the Structure of the Self* spans too many different areas of specialization for any one person to review it. So it would seem that the one person who wrote it actually must have comprised several different ghosts in the machine, each ghostwriting a different chapter. Or perhaps she is in reality just an oversized Swiss army knife, presuming to dissect any fodder on the chopping block.

> provisional members, and interlopers tend to rank among the lowest subordinates of all. Accordingly, the more they diverge – in thought, appearance or pedigree – from the tradition, the closer to the bottom of the hierarchy they are likely to be found, and the more blatant the exercises of power that keep them there.[3]

But as of that writing, I had not yet been gently eased out of the United States, nor gently eased out of my tenured full professorship, nor gently eased out of my retirement benefits, nor gently eased out of my agreement with CUP, nor gently eased out of any remaining status in that professional hierarchy. This gentle and easy sequence of events attests to the predictive power of the theory developed in Chapter I, legitimates its aspiration to truth, and secures my role as experimental guinea pig of my own theory. For that reason, among others, I have made no revisions of content, aside from minor corrective line-edits, in the main text of this second edition. Perhaps the passage of time will gradually disclose the predictive power of theories developed in subsequent chapters of the project as well.

Socrates reminds us that a hierarchy of status is not the same as a hierarchy of quality. I recurred to this useful advice each time I was forced to choose between them, by refusing repeatedly, under institutional pressure, to publish *Rationality and the Structure of the Self* prematurely or in butchered form. I have never regretted my decision to pay any price necessary in order to publish this work at the highest standard of philosophical achievement of which I am capable. Of course the price of doing the very best philosophical work I had it in me to do should not have been that expensive. But it has been more than repaid by the insights it has yielded into the *de facto* workings of the profession.

The most important of these insights may be worth sharing: Whether your work is blacklisted, ignored, or simply overlooked by your colleagues does not necessarily undermine, and may even aid and abet your ability to produce the best work you possibly can. If you are lucky enough to have access to a laptop and a library,[4] no one can stop you from doing that work unless you let them. Document and archive it properly, and you will get your 15 minutes eventually. We all do. Write for that audience, not this one.

These insights have yielded a freedom to say and do and write what I want that my previous investment in the institutional hierarchy of academic

---

[3] *Rationality and the Structure of the Self*, Chapter I. General Introduction to the Project: The Enterprise of Socratic Metaethics, 21 (both volumes), below.

[4] – and perhaps some sense of kinship with the many artists who choose to moonlight alongside day jobs that pay the bills. Philosophy is much cheaper to finance and just as easy to feed. *Teaching* philosophy of course should be much more than that. But if its proffered working conditions effectively thwart any such activity worth the name, then it is much less; and may offer much less food for thought than other available day jobs such as managing an office or driving a cab.

philosophy had not returned. Having been gently eased out of the profession, I can now indulge without guilt the luxury of devoting myself to the discipline; and of doing even more of the very best work I possibly can, regardless of whom it offends.[5] The anonymous acclaim collected at the back of these volumes lends empirical support to these insights, while minimizing the professional dangers that public exposure of the culprits would bring. My choices have turned me into a walking institutional critique; and I find I enjoy this new persona very much.[6]

Recently a very eminent colleague of long standing, almost exactly my age and the recipient of a named chair at a top-ranked university, invited me to lunch and inquired as to how things were going with *Rationality and the Structure of the Self*. I reported to him what I have reported to other curious bystanders, and what I have now reported here, once and for all, in this Preface. He questioned whether "getting kicked out of the field" was an accurate description of my experience. He inquired into the events and personalities at the academic institution that had delivered the boot. And he described with relish his review of another mutual colleague's recent two-volume work. He offered to send me both the review, published a year after the appearance of both volumes, and the volumes themselves. I appreciated the opportunity to make some useful comparisons. My colleague had read the text carefully, annotating key passages in the margins and indexing them on the flyleaves. His review was fair, thorough, attentive to the argument, and appropriately respectful of the author's diligent efforts and exalted professional status. It regretfully concluded that the work under review was deeply misguided and historically worthless.

My efforts in *Rationality and the Structure and the Self* were only slightly less diligent (a measly 1,212 total printed pages for my two volumes to 1,365 for his). But my professional status is considerably less exalted. In fact, it is so microscopically Tom Thumb-diminutive that *Rationality and the Structure of the Self* offers no professional incentive whatsoever, aside from unattributed use of its ideas, to read it. There is no legitimate professional end to which attention to this project is a means. Neither academic standing, nor peer recognition, nor professorial approval, nor enhanced professional connections, nor powerful patronage, nor job offers, nor tenure, nor journal publication, nor external research funding, nor any other professional rewards will accrue for publicly disclosing one's acquaintance with or interest in this work. Indeed, any such attention spent must debit and justify the time,

[5] After all, what are any offended parties going to do about it? Kick me out of the field?

[6] However, I am no match for Gene Roddenberry's Borg, the uncontested winners of the Pink Floyd Lifetime Achievement Award for institutional critique in the peripatetic tradition.

attention and energy thereby lost to other endeavors more conducive to professional flourishing.

As for its worth, the only reasons to read *Rationality and the Structure of the Self* (in private, of course, or else concealed in a plain brown paper bag) are stubborn curiosity about that very question: Was it, in fact, really worth it? – plus whatever historical worth its curiouser and curiouser history has inadvertently conferred. I am glad it has caught the attention of the curious, and I value their curiosity.

I hope your curiosity will be slaked by what you find in the following pages; that they will answer that question, both to your satisfaction and to mine; and that the answer you find there will have been worth the trouble of seeking it out.

Adrian M. S. Piper
Berlin, 24 January 2013

It is much more honorable and much easier
not to suppress others, but to make yourselves as good as you can.[7]

[7] Plato, *Apology XXX*, in *Euthyphro, Apology, Crito*, Trans. F. J. Church and Robert D. Cumming (New York: Bobbs-Merrill, 1956)

Acknowledgements to Volume I

My first inkling that there was something amiss with the Humean conception of the self came before I knew enough Western philosophy to call it that. I am grateful to Allen Ginsberg, Timothy Leary, Edward Sullivan and Swami Vishnudevananda for urging me to read the *Upanishads*, *Bhagavad Gita* and *Yoga Sutras* in 1965. I am grateful most of all to Phillip Zohn for his willingness to argue with me at length about the import of these texts, and for introducing me to Kant's *Critique of Pure Reason* in 1969, after reading an art text of mine on space and time ("Hypothesis") that inadvertently echoed its doctrine of transcendental idealism. The influence of all of these works on my thinking has informed my (you will pardon the pun) critical and skeptical approach to the Humean conception from the beginning.

This project has been in production for a very long time. The ancestor of the concept of pseudorationality introduced in Chapter VII of Volume II was my undergraduate Social Sciences Phi Beta Kappa Medal Honors Thesis, "Deception and Self-Deception" (City College of New York, 1974). I am grateful to Martin Tamny, Arthur Collins and David Weissman for their guidance and input at that stage. The ancestor of the analysis of cyclical and genuine preference in Chapter IV of Volume I and Chapter III of Volume II was Chapter II of my Second-Year Paper, "A Theory of Rational Agency" (Harvard University, 1976), for advice and comments on which I am indebted to John Rawls. Both ancestors liased in revised form in my dissertation, "A New Model of Rationality" (Harvard University, 1981) under John Rawls and Roderick Firth, in whose debt I permanently remain. Professor Firth provided the sounding board, the detailed and rigorous criticism, and the personal encouragement that has helped preserve my faith in the value of this project. I am deeply grateful for his involvement with it, and to have known him as a teacher and colleague.

My animated discussions with Professor John Rawls, both about my work and about the role of the utility-maximizing model in his work, were absolutely crucial to my conviction that I was on to something. His example as a scholar and teacher, the breadth and depth of his learning, and his magisterial achievement in *A Theory of Justice* have remained an inspiration to me in all of my work. I rank Rawls's achievement as a *theory-builder* – a philosopher who constructs substantive theories – with those of the middle and late Plato, Aristotle, Hobbes, Kant, and Habermas. A *critic*, by contrast, is a philosopher who mostly criticizes, improves upon, or demolishes theory-builders' theories. The quintessential critic would be the slice-'em-and-dice-'em Socrates of the early Platonic dialogues. But some might also count St. Thomas Aquinas, Sidgwick, the later Wittgenstein, and Ryle among the philosophical critics, for different reasons. Philosophers may reasonably disagree about how some of these examples are to be classified, and most

philosophers evince both theory-building and critical inclinations to varying degrees. But the distinction is nevertheless useful, because training in analytic philosophy is by default training in how to be a critic: We study the views of famous philosophers, learn how to detect areas of inconsistency or fault or lack, and then learn how to correct, supplement or level them. There is no way to teach theory-building, except by encouraging students to have confidence in their intuitions. So if we happen to incline toward theory-building, we are pretty much on our own, because there are no ground rules about how to proceed. In developing the theory defended in this project, I was fortunate from the very beginning to receive good advice about how to proceed, from another theory-builder who had already been there and done that. The ground rules Rawls taught me were three:

> (1) Anchor your theory in relation to identifiable current problem(s) or controversies. Describe the problems, analyze some recent arguments that purport to solve them, and explain the ways in which these arguments fail. Then briefly sketch how your theory avoids these failures, so that your readers will be able to locate your theory on their own map of philosophical issues in a way that confers meaning and importance on it for them.
>
> (2) Anchor your theory relative to the views, with which you disagree, of other philosophers who have worked on the problem and have received attention for their efforts. Discuss those views, explain what's wrong with them, and describe how your theory avoids the criticisms you make of their views. Refer to these opposing views in developing your own, in order to bring your theory into connection with a larger, ongoing philosophical discussion among your peers.
>
> (3) Avoid cooking up a straw man to attack. Show that you take your opponents' views seriously, by making the best and most sympathetic case for them you possibly can, before showing how they disappoint despite your best efforts. The worst that can happen is that really understanding your opponents' views will convince you to modify your own.

In this project I have tried to honor Rawls's ground rules as best I can, in order to honor him as my teacher and their author, and also all of those others from whom I have learned so much by disputing their views in the following pages.

I have also benefited by teaching and discussing extensive portions of both volumes of this project with several generations of graduate students at the University of Michigan, Stanford, Georgetown and USCD – particularly Richard Dees, Jeffrey Kahn, Brian Leiter, Alan Madry, Minerva San Juan McGraw, David Reed-Maxfield, Joel Richeimer, Laura Shanner, Cristel

Steinvorth, and Sigrun Svavarsdottir; and fifteen years' worth of brilliant and feisty undergraduates at Wellesley College.

Chapter I of both volumes, "General Introduction to the Project: The Enterprise of Socratic Metaethics," was drafted during an unpaid leave of absence from Wellesley College during early 1998 and funded by an NEH College Teachers' Research Fellowship. The NEH support came at a crucial moment and I am deeply grateful for it. This chapter incorporates and modifies some passages and sections of my "Two Conceptions of the Self," published in *Philosophical Studies 48*, 2 (September l985), 173-197 and reprinted in *The Philosopher's Annual VIII* (1985), 222-246. The discussion of Anglo-American philosophical practice that appears in Sections I.2 and I.3 benefited from comments by Anita Allen, Houston Baker, Paul Boghossian, Ann Congleton, Joyce Carol Oates, Ruth Anna Putnam and Kenneth Winkler, as well as by members of the audience to the 1994 Greater Philadelphia Philosophy Consortium symposium, "Philosophy as Performance" at which these remarks were originally presented. The chapter received its near-final form during my tenure as a Research Scholar at the Getty Research Institute during the academic years 1998-1999. For providing me with all of the conditions I requested – some very idiosyncratic – as necessary for me to make substantial progress on this and many other parts of this project, my gratitude to the Institute knows no bounds. My debt of thanks to Brian Davis, Larry Hertzberg, Karen Joseph, Michael Roth, and Sabine Schlosser is particularly great. While there I also benefited a great deal from discussion of these and related topics with Reinhart Meyer-Kalkus. I would also like to thank Naomi Zack for her interest and willingness to publish an earlier version of this chapter, despite its length, in her edited collection, *Women of Color and Philosophy* (New York: Blackwell, 2000).

Chapter II, "The Belief-Desire Model of Motivation," was first drafted in 1981, while I was an Assistant Professor at the University of Michigan. I learned much from discussing the issues with Richard Brandt, William Frankena, Allan Gibbard, Jaegwon Kim, David Velleman, Nicholas White, and Stephen White, however much we in the end agreed to disagree. The chapter was redrafted in 1985, after having spent two wonderful and productive years at the Stanford University Philosophy Department on an Andrew Mellon Post-Doctoral Fellowship from 1982 to 1984. While there I benefited from discussing action theory with Michael Bratman and philosophy of science with John Duprés. Not until my year at the Getty Research Institute in 1998 was I able to return to this part of the project. The enthusiasm and dedication of the Getty staff in putting at my disposal all of the research and administrative assistance I needed, and more, to update and revise it in light of more recent discussions helped me to believe in the importance of doing so.

Work on Chapters III and IV was partially supported by the Mellon Post-Doctoral Fellowship, a Georgetown University Faculty Research Grant in

1988, and a Woodrow Wilson International Scholars' Fellowship in 1988-1989. At Georgetown I profited from discussions with Wayne Davis, Terry Pinkard and Henry Richardson. I also spent many, very fruitful hours discussing this material with colleagues at the University of Michigan, all Humeans to a man. To them I am most grateful of all for pulling no punches in their attempts to dissuade me from my views, from these chapters, and not least of all from this project. Had they not put those views to the test by resorting to every possible tactic of dissuasion, I would have had no proof that my views could withstand them. To have that proof – to know that my philosophical position was able to survive the gauntlets devised by some of the very best minds in the field – is the invaluable gift that I owe to them. I should particularly like to thank Allan Gibbard and David Velleman for conversation. Many other individuals have helped me in the writing of these two chapters, including Glenn Loury, Michael Slote, Robert Audi, David Levy, and especially Ned McClennen for extensive comments on earlier drafts. I was honored by the opportunity to present both to a group of trained economists at the Economics and Rhetoric Seminar, held at the Academia Vitae in Deventer, The Netherlands, in June 2006. I am particularly grateful to Arjo Klamer, Dierdre McCloskey and P. W. Zuidhof for beneficial discussion that has improved their final form.

I was helped by discussion of the first draft of Chapter V at a University of Michigan Faculty Seminar, and particularly by comments from Richard Brandt, Arthur Burks, Allan Gibbard, Louis Loeb, Peter Railton, Nicholas White, and Stephen White. An earlier version was published under the title, "A Distinction Without a Difference," in *Midwest Studies in Philosophy VII: Social and Political Philosophy* (1982), 403-435. Chapter VI was completed and delivered to the Scholars' Seminar at the Getty Research Institute in November 1998, and to the Philosophy Department at the University of Minnesota in October 1999. I am grateful to both audiences for constructive comments and suggestions for improvement. Chapter VII is the outcome of over two decades of intense and satisfying – and, aside from Paul Coppock's insightful comments, largely solitary – labor on Thomas Nagel's *The Possibility of Altruism*. This part of the project taught me much more about patience and persistence than I ever could have expected when, after completing the chapter to my satisfaction a first time, I then allowed it to be irretrievably misplaced and had to reconstitute it from scratch (a scholar's worst nightmare in the pre-computer era). I am deeply grateful to the Woodrow Wilson International Research Center of the Smithsonian Institution for extending my International Scholars' Fellowship for a second year, 1989-90, so that I could do this.

Earlier versions of parts of Chapter VIII were published under the following titles: "Two Conceptions of the Self," *Philosophical Studies 48*, 2 (September l985), 173-197, reprinted in *The Philosopher's Annual VIII* (1985), 222-246; "Moral Theory and Moral Alienation," *The Journal of Philosophy*

LXXXIV, 2 (February 1987), 102-118; and "Michael Slote's *Goods and Virtues,*" reviewed for *The Journal of Philosophy LXXXIII*, 8 (August 1986), 468-73. Work on this chapter was supported by the Mellon Fellowship. Earlier versions of the discussions of Frankfurt and Watson were presented to the Philosophy and Anthropology Group and the Department of Philosophy, both at the University of Michigan; and the Departments of Philosophy at Stanford, U. C. Berkeley, the University of Minnesota, and the University of Pennsylvania. I learned much from comments received on those occasions, and from detailed criticism and feedback by Michael Bratman, Jeffrey Evans and Allan Gibbard. I am equally grateful to Akeel Bilgrami, Jeffrey Evans and members of the Philosophy Department audiences at Wayne State University, Penn State, Georgetown, the University of California at San Diego, North Carolina State, Wesleyan, Memphis State, and the University of Minnesota for comments and criticism of my discussion of Williams.

Section 1 of Chapter IX was delivered in a slightly different form to the American Philosophical Association Pacific Division Convention in March 1995 in an Author Meets Critics session on Elizabeth Anderson's *Value in Ethics and Economics*; and later published under the title, "Making Sense of Value," in *Ethics 106*, 2 (April 1996), 525-537. An earlier version of Section 4 was delivered to the Moral Philosophy Colloquium at the American Philosophical Association Pacific Division Convention, Los Angeles, California in March 1986; and published under the title, "Instrumentalism, Objectivity, and Moral Justification," in *American Philosophical Quarterly 23,* 4 (October 1986), 373-381.

Chapter X originated in September 1976 as a paper, "Continuing Persons and the Original Position," for John Rawls's graduate seminar in Moral Psychology, and I am grateful for his comments on it. A lecture by Joshua Cohen on Social Contract Theory in the Fall of 1978 at MIT had a salutary effect on Section IX.2. I have also benefited from criticisms of an earlier draft of this chapter by Peter Dalton. Parts were published under the title, "Personal Continuity and Instrumental Rationality in Rawls' Theory of Justice," in *Social Theory and Practice 13*, 1 (Spring 1987), 49-76. Work on this chapter was supported by a University of Michigan Rackham Faculty Fellowship and the Mellon Fellowship. The final draft was completed during my year at the Getty, as was the final draft of Chapter XI. Chapter XII, originally my term paper for John Rawls's Moral and Political Philosophy course at Harvard in the Spring of 1975, was also revised and completed during my wonderful and productive year at the Getty. I am grateful to Rawls, David Auerbach and Warner Wick for helpful criticisms of earlier drafts. An earlier version was published under the title, "Utility, Publicity and Manipulation," in *Ethics 88*, 3 (April 1978), 189-206.

For criticisms of an earlier draft of Chapter XIII I would like to thank Anita Allen, Annette Baier, Margaret Carroll, John Deigh, Michael Stocker, and Judith Jarvis Thomson. A protodraft of Chapter XIV originally formed the

Appendix to my dissertation. I am grateful to Rawls for persuading me of the importance of dealing with Hume straight off, and for his criticisms and encouragement throughout. I also would like to thank Marcia Baron for comments on an earlier draft of this chapter. A more recent version was published under the title, "Hume on Rational Final Ends," in *Philosophy Research Archives XIV* (1988-89), 193-228. Robert Audi's comments and criticisms, and that of an unidentified referee, improved that version immensely. Chapter XV is entirely my own doing, and I have no one to blame but myself.

There is no way for me to express my gratitude and indebtedness to the very few individuals who provided encouragement and support during the final stretch of time in which I brought this project to completion. During two years of unpaid and extremely stressful medical leave from Wellesley College from Winter 2001 to Fall 2002, Bill Cain, Joe Feagin, Terry Irwin, Mark Kaplan, James Kodera, Ruth Barcan Marcus, Julie Matthaei, Reinhart Meyer-Kalkus, Susan Neiman, Robert Rubinowitz, Stephen Schiffer, Hedwig Saxenhuber, Georg Schöllhammer, Ann Stephens, and Joan Weiner extended themselves beyond the bounds of collegial or moral obligation by letting me know, each in their own way, the importance and value to them that I do so. Their encouragement was crucial. My debt to Ruth Barcan Marcus for her steadfast friendship is beyond measure. The research and administrative help provided, under less than ideal conditions and great generosity of spirit, by Robert Del Principe was invaluable. His patience, resourcefulness, persistence and good humor in obtaining the sources I needed under the most stressful conditions, and tolerating without complaint twelve years' worth of my unending incipient hysteria has manifested both heroism and martyrdom of the highest order. My debt to him is incalculable. Without the moral support of all of these good people this project would not have been possible. The final draft was begun under conditions of extreme personal hardship, in virtually complete solitude during the long, hot summer of 2003; and received its final form in the sheltering anonymity and safety of the city of Berlin in early 2008. I am profoundly grateful that it is there, and that I am there. For the unique opportunity to live and test the values defended in this project, I would like to thank the faculty and administration of Wellesley College; I commend this work in exile to them. For the strength, the solace and the sanctuary I have been blessed to find in reading, writing and teaching philosophy I am grateful most of all.

Chapter I. General Introduction to the Project: The Enterprise of Socratic Metaethics

Buffeted and bruised by the currents of desire and longing for once to ride the wave, we may cast about for some buoyant device from which to chart a rational course; and, finding none, ask ourselves these questions:

> Do we at least have the *capacity* ever to do anything beyond what is comfortable, convenient, profitable, or gratifying?
>
> Can our conscious explanations for what we do ever be anything more than opportunistic *ex post facto* rationalizations for satisfying these familiar egocentric desires?
>
> If so, are we capable of distinguishing in ourselves those moments when we are in fact heeding the requirements of rationality, from those when we are merely rationalizing the temptations of opportunity?

I am cautiously optimistic about the existence of a buoyant device – namely reason itself – that offers encouraging answers to all three questions. Without hard-wired, principled rational dispositions – to consistency, coherence, impartiality, impersonality, intellectual discrimination, foresight, deliberation, self-reflection, and self-control – that enable us to transcend the overwhelming attractions of comfort, convenience, profit, gratification … and self-deception, we would be incapable of acting even on these lesser motives. Or so I argue in this project. I take it as my main task to spell out in detail the ways in which these hard-wired, principled dispositions rationally structure the self; in effect, outfit human beings with high-caliber cognitive equipment we are not yet able to fully exploit.

This task thus depends on a distinction between two different but related aspects of rationality. I describe as *egocentric rationality* action guided by considerations of comfort, convenience, profit, or gratification – in short, by principles spelled out in what I call the *Humean conception of the self*. In Volume I, I define, dissect and criticize in detail this desire-centered conception as formulated in late-twentieth century Anglo-American analytic philosophy. Chapter VI of Volume I defends the claim that "egocentric" is the correct description of this conception, against objections from its advocates. Although Volume I very often catalogues the shortcomings of this widely held view, it ultimately argues that the strengths of the Humean conception can be fully exploited only by situating it as a special case within a larger context.

This larger context is given by principles of what I call *transpersonal rationality*, i.e. principles governing the hard-wired rational dispositions listed above. In Volume II, I analyze these principles as constitutive of what I call the *Kantian conception of the self*. I describe these principles as "transpersonal" because they direct our attention beyond the preoccupations and interests of the ego-self, including its particular, defining set of moral and theoretical

convictions; and apply in equal measure to oneself and others. Transpersonal principles thus often require us to transcend considerations – even principled considerations – of personal comfort, convenience, profit, or gratification, whether acting on our own behalf or on behalf of another. Chapter VIII of Volume I contains discussion of the more familiar notions of impersonal and impartial principles, which each relate to transpersonal principles as instance to concept. Chapter V of Volume II contains an extended account of what it would be like for us to guide all of our behavior by transpersonal principles, whether self- or other-directed; and Chapters VII through XI an account of how and why we compulsively try but usually fail to do so.

Thus my distinction between transpersonal and egocentric rationality cuts across the traditional distinction between theoretical and practical reason. Transpersonal principles include so-called theoretical ones of coherence and logical consistency, as well as so-called practical principles of foresight and self-control. Similarly, egocentric principles may include so-called theoretical ones relating cause to effect of the sort that are to be found in Machiavelli, as well as so-called practical principles that govern the maximization of personal gratification. I use the slightly pejorative locution "so-called," because I believe that this distinction has been made to carry much more weight than it can bear, *pace* Kant, and in the end does not come to much. In Volume II I defend this opinion at length.

Sections 1 through 6, following, of this General Introduction to the Project elaborate the intuitive distinction between egocentric and transpersonal rationality through its application to the particular case that most personally motivates this project for me, and that I hope will also motivate the reader to patiently but persistently follow its single line of argument through two large volumes, one section at a time. That particular case is current philosophical practice itself. I choose to discuss this case, first, because it is the one that most urgently compels me to address the three questions with which I began this Introduction; and second, because I do not find widespread recognition in the field that philosophers' virtually universal obsession with the topic of rationality – with defining it, critiquing it, defending it, rejecting it, elaborating alternatives to it – is implicitly an activity of professional *self*-definition, *self*-critique, *self*-defense, *self*-rejection, and *self*-elaboration of the methodological foundations on which the practice of philosophy itself rests. The resulting failure to apply self-consciously to the practice of philosophy the principles of rationality that philosophy itself champions has bad consequences both for theory and for practice; and, I believe, leads us to underestimate the necessity of clarifying in what our actual relation to rationality consists, even as we continue to be obsessed by it. By directing the above three questions in the first instance specifically to philosophical practice, I hope to find consensus among philosopher-readers of this Introduction on the importance of trying self-consciously to answer

them, even if not on the importance of the particular answers I myself offer in this project. I recur often to this particular test case in the two-volume argument that follows.

1. Transpersonal Rationality and Power

In order to actualize the potential for transpersonal rationality, one must first genuinely value it. That is, one must value both rational behavior that transcends the personal and egocentric, and also the character dispositions which that behavior expresses. According to Nietzsche, the capacity for reason becomes a value when it is valorized by a "slave morality" that assigns highest priority to the character dispositions of transpersonal rationality and the spirit at the expense of natural human instincts. Like a good *Untertan*, I intend to do exactly that in this project: not argue for the value of transpersonal rationality, but rather presuppose its value, and argue for our innate ability to turn it into a fact – what Kant optimistically calls the fact of reason.

Thus I am going to presuppose that if a person's freedom to act on her impulses and gratify her desires is constrained by the existence of equally or more powerful others' conflicting impulses and desires, then she will need the character dispositions of transpersonal rationality to survive; and will assign them value accordingly. The more circumscribed her freedom and power, the more essential to survival and flourishing the character dispositions of transpersonal rationality become. And to the extent that such a person's power to achieve her ends is limited by a distribution of scarce social or material resources often less than fair or favorable to herself, she will to that extent, at least, value the character dispositions of transpersonal rationality as a needed source of strength and solace. Genuinely valuing the capacity for reason, then, proceeds from concrete experience of its power.

On these assumptions, the valorization of the character dispositions of transpersonal rationality that typify a "slave morality" does not express mere sour grapes, as Nietzsche sometimes suggests in his more contemptuous moments. Nor does it merely make a virtue of necessity, although it does at least do that. It recognizes an intrinsic good whose value may be less evident to those for whom it is less necessary as an instrument of survival:

> How long will you wait to think yourself worthy of the highest and transgress in nothing the clear pronouncement of reason? ... Therefore resolve before it is too late to live as one who is mature and proficient, and let all that seems best to you be a law that you cannot transgress. ... This was how Socrates attained perfection, attending to nothing but

> reason in all that he encountered. And if you are not yet Socrates, yet you ought to live as one who would wish to be a Socrates.[1]

Think of these injunctions as conjointly constitutive of the *Socratic ideal*. As the product of biographical fact, Epictetus' loyalty to the Socratic ideal, and in particular his injunctions to "transgress in nothing the clear pronouncement of reason," and to "atten[d] to nothing but reason in all that [we] encounte[r]" are an expression of wisdom borne of the personal experience of enslavement. They attest to the valuation and cultivation of transpersonal rationality as the weapon of choice for the unempowered to use on their own behalf. They both underwrite Nietzsche's analysis of reason and the spirit as central values of a "slave morality," and demonstrate how that "slave morality" may have a kind of dignity that *übermenschlichen* views lack.

For if a person's freedom and power to gratify his impulses is greater, then he may well find the egocentric indulgence of emotion, spontaneity, instinct, and the manipulation of power more attractive; and development of the character dispositions of transpersonal rationality correspondingly less necessary, interesting, or valuable. After all, such individuals have at hand other reserves – of wealth, status, influence and coercion – on which to draw to achieve their ends. The unique quality of ends that the character dispositions of transpersonal rationality themselves inspire therefore may be accorded correspondingly less importance, if they are noticed in the first place. For such individuals, the Socratic ideal is no ideal at all; and perfunctory lip service to the value of rational decision-making is merely one dispensable strategy among others for facilitating the ongoing indulgence of impulse.

Philosophy as an intellectual discipline is fundamentally defined and distinguished from other intellectual disciplines by its *de facto* loyalty to the character dispositions of transpersonal rationality, and so to the Socratic ideal. Anglo-American analytic philosophy is committed to these values with a particularly high degree of self-consciousness. Whatever the content of the philosophical view in question, the norms of transpersonal rationality define its standards of philosophical exposition: clarity, structure, coherence, consistency, subtlety of intellectual discrimination. And as a professional and pedagogical practice, philosophy is ideally defined by its adherence to the norms of rational discourse and criticism. In philosophy the appeal is to the other's rationality, irrespective of her personal, emotional or professional investments, with the purpose of convincing her of the veracity of one's own

[1]Epictetus, *Enchiridion* LI. I have consulted two translations: P.E. Matheson (Oxford: Clarendon Press), reprinted in Jason L. Saunders, Ed. *Greek and Roman Philosophy after Aristotle* (New York: The Free Press, 1966), 147; and George Long (Chicago: Henry Regnery Co., 1956), 202-203.

point of view. It is presumed that this purpose has been achieved if the other's subsequent behavior changes accordingly.

This presumption is fueled by philosophy's unsupervised influence in the political sphere – of Rousseau on the French Revolution, Locke on the American Revolution, Marx on Communism, Nietzsche on the Second World War, Rawls's Difference Principle on Reaganomics. In the private and social sphere, rational analysis and dialogue may just as easily give way to unsupervised imbalances in power and freedom, paternalistic or coercive relationships, or exploitative transactions. But even here it is not impossible for philosophy to have its influence: in turning another aside from an unethical or imprudent course of action, or requiring him to revise his views in light of certain objections, or altering his attitudes toward oneself, or influencing others to accommodate the importance of certain philosophical considerations through compromise, tolerance, or mutual agreement.

In both spheres, then, the attempt rationally to persuade and to conduct oneself rationally toward others is an expression of respect, not only for their rational capacity, but thereby for the alternative resources of power – coercion, bribery, retaliation, influence – they are perceived as free to use in its stead. Toward one who is perceived to lack these alternative resources, no such respect need be shown, and raw power may be displayed and exercised more freely, without the limiting constraints of rational justification. For, as Hobbes reminds us,

> [h]onourable is whatsoever possession, action, or quality, is an argument or sign of power. ... And therefore to be honoured, loved, or feared of many, is honourable; as arguments of power. ... To speak to another with consideration, to appear before him with decency, and humility, is to honour him; as signs of fear to offend. To speak to him rashly, to do any thing before him obscenely, slovenly, impudently, is to dishonour.[2]

Hobbes is wrong to think that treating another with respect is nothing but an expression of fear of the other's power. But he is surely right to think that it is at least that. On Nietzsche's refinement of Hobbes' analysis, the appeal to reason expresses respect for another's rational autonomy to just and only that extent to which it simultaneously expresses fear of the alternative, nonrational ways in which that autonomy may be exercised. On Nietzsche's analysis of rational conduct, Hobbes and Kant may both be right.

So philosophy's traditional commitment to the Socratic ideal is one quintessential expression of a "slave morality" that acknowledges the danger of unrestrained instinct and the egocentric use of power in its service, by to varying degrees constraining and sublimating instinct, impulse, and the manipulation of power into a rational exercise of intellect and will that brings its own fulfillments:

---

[2]Thomas Hobbes, *Leviathan*, Ed. Michael Oakeshott (New York: Collier, 1977), 75, 74.

> The ignorant man's position and character is this: he never looks to himself for benefit or harm, but to the world outside him. The philosopher's position and character is that he always looks to himself for benefit and harm. The signs of one who is making progress are: he blames none, praises none, complains of none, accuses none, never speaks of himself as if he were somebody, or as if he knew anything. When he is hindered, he blames himself. ... He has got rid of desire, and his aversion is directed no longer to what is beyond our power [i.e. the body, property, reputation, office, and, in a word, everything that is not our own doing] but only to what is in our power [i.e. thought, impulse, desire, aversion, and, in a word, everything that is our own doing] and contrary to nature. In all things he exercises his will temperately.[3]

The philosopher, according to Epictetus, foregoes the egocentric gratification of desire and acquisition of external goods and power for the sake of cultivating the character dispositions of transpersonal rationality. Seeing that these two alternatives frequently conflict, she "atten[ds] to nothing but reason in all that [she] encounter[s]." The centrality and universality of the character dispositions of transpersonal rationality to the discipline of philosophy, enduring over nineteen centuries, may explain why almost all philosophers, regardless of their express philosophical views on the value of rationality, try to muster the resources of rational argumentation, analysis, and criticism to defend those views. The consistency and sincerity with which they try to live up to the Socratic ideal bespeaks the seriousness of their intent to avoid the dormant alternatives.

2. Transpersonal Rationality as Philosophical Virtue

The priority accorded to the character dispositions of transpersonal rationality in the practice of philosophy receives a more contemporary formulation in the following Anglo-American analytic version of the Socratic ideal:

> [G. E.] Moore ... invented and propagated a style of philosophical talking which has become one of the most useful and attractive models of rationality that we have, and which is still a prop to liberal values, having penetrated far beyond philosophical circles and far beyond Bloomsbury circles; it is also a source of continuing enjoyment, once one has acquired the habit among friends who have a passion for slow argument on both abstract and personal topics. When I look back to the Thirties and call on memories, it even seems that Moore invented a new moral virtue, a virtue of high civilization admittedly, which has its ancestor in Socrates' famous following of an argument wherever it may lead, but still with a quite distinctive modern and Moorean accent. Open-

[3] *op. cit.* Note 1, XLVIII; also see I.

> mindedness in discussion is to be associated with extreme literal clarity, with no rhetoric and the least possible use of metaphor, with an avoidance of technical terms wherever possible, and with extreme patience in step-by-step unfolding of the reasons that support any assertion made, together with all the qualifications that need to be added to preserve literal truth, however commonplace and disappointing the outcome. It is a style and a discipline that wring philosophical insights from the English language, pressed hard and repeatedly; as far as I know, the style has no counterpart in French or German. As Nietzsche suggested, cultivated caution and modesty in assertion are incompatible with the bold egotism of most German philosophy after Kant. This style of talking, particularly when applied to emotionally charged personal issues, was a gift to the world, not only to Bloomsbury, and it is still useful a long way from Cambridge.[4]

The writer is Stuart Hampshire, and in this passage he describes as an historical fact a more recent ideal of philosophical practice that speaks to some of the motives and impulses that attract many into the field. The essence of the ideal remains Socratic: clarity and truth as a goal, with patience, persistence, precision, and a nonjudgmental openness to discussion and contention as the means.

Hampshire is right to describe this ideal as a "new moral virtue ... of high civilization." It is a moral virtue because it imposes on one the obligation to subordinate the egocentric desires to prevail in argument, to shine in conversation, or to one-up one's opponent to the disinterested ethical requirements of impartiality, objectivity and transpersonal rationality in discussion. And it is a virtue of high civilization because it is not possible to achieve this virtue – or even to recognize it as a virtue – without already having cultivated and brought to fruition certain civilized dispositions of character, tastes and values that override the desire to prevail. Thus this moral virtue stands at the very center of a "slave morality" that sublimates the desire to prevail to the imperatives of reason and the spirit. These imperatives, in turn, find expression in what Mill calls the higher pleasures of the intellect and moral and aesthetic sensibility. They presuppose the victory of "slave morality" in subjugating instinct and the egocentric exercise of power to the rule of reason and its attendant ethical values of fairness and impartiality in thought and action. This virtue of high civilization, then, presupposes both its participants' transpersonal rationality and also their achievement of a mutually equitable balance of power – however the material and social instruments of power may be distributed.

---

[4]Stuart Hampshire, "Liberator, Up to a Point," *The New York Review of Books XXXIV*, 5 (March 26, 1987), 37-39.

Thus this ideal can have meaning only for someone for whom basic psychological and spiritual needs for self-worth, and moral needs for the affirmation of self-rectitude are not so pressing that every dialectical encounter with others – whether written or conversational – is mined for its potential to satisfy them. So when we say of such a person that he is civilized, we may mean, among other things, that in conversation he is disposed to be generous in according credibility to his opponent's view, gracious in acknowledging its significance, patient in drawing forth its implications, and graceful in accepting its criticism of his own. Someone who has mastered this new moral virtue of high civilization is someone for whom philosophical practice expresses an ideal of personal *civility*; a civility made possible only by the control and sublimation of instinct, impulse, desire, and emotion.

The higher pleasure of doing philosophy in the style Hampshire describes is then the disinterested pleasure of thinking, considering, learning and knowing as ends in themselves, and of giving these pleasures to and receiving them from others involved in the same enterprise, in acts of communication. Plato was surely right to suggest that we are driven to seek erotic pleasure from others by the futile desire to merge, to become one with them. Erotic desire is ultimately futile for reasons of simple physics: we are each stuck in our own physical bodies, and you cannot achieve the desired unity by knocking two separate physical entities together, no matter how closely and repeatedly, and no matter how much fun it is to do the knocking.

Intellectual unity with another is a different matter altogether, however; and the kind Hampshire describes is particularly satisfying because it does not require either partner to submerge or abnegate herself in the will or convictions of the other. It does not require sharing the same opinions, or suppressing one's own worldview, or deferring or genuflecting to the other in order to achieve agreement with him. Rather, the enterprise is a collaborative one between equals who pool their philosophical resources. By contributing questions, amendments, refinements, criticisms, objections, examples, counterexamples, or elaborations in response to the other's philosophical assertions, we each extend and enrich both of our philosophical imaginations past their individual limits and into the other's domain. There are few intellectual pleasures more intense than the *Aha-Erlebnis* of finally understanding, after long and careful dialogue, what another person actually means – unless it is that of being understood oneself in this way.

The ground rules for succeeding in this enterprise are ethical ones. By making such assertions as clearly as I can, I extend to you an invitation to intellectual engagement; and I express trust, vulnerability and respect for your opinion in performing that act. I thereby challenge you to exercise your trained philosophical character dispositions – for impartiality, objectivity, and hence transpersonal rationality – in examining my assertions; and to demonstrate your mastery of the enterprise in the act of engaging in it. This is

the challenge to perform, in the practice of dialogue and conversation, at the ethical level made possible by our basic human capacities for language, logic and abstraction; and to bring those capacities themselves under the purview and guidance of our conception of right conduct. By engaging in the enterprise of philosophical dialogue, we challenge each other to observe the ethical and intellectual obligations of philosophical practice.

In this enterprise, I have failed if you feel crestfallen at having to concede a point, rather than inspired to elaborate upon it; or ashamed at having missed a point, rather than driven to persist in untangling it; or self-important for having made a point, rather than keen to test its soundness. After all, the goal of the enterprise is to inspire both of us with the force of the ideas we are examining, not to make either of us feel unequal to considering them, or smug for having introduced them. Too often we conceive of moral virtue as having to do only with such things as helping the needy, keeping promises, or loyalty in friendship – as though performing well in these areas relieved us of the obligation to refrain from making another person feel stupid, ashamed or crazy for voicing her thoughts; or ourselves feel superior for undermining them. When teachers fail to impart a love of philosophy to their undergraduate students, or drive graduate students, traumatized, out of their classes and out of the field, it is often because these elemental guidelines for conducting the enterprise – guidelines that express the simple truth that a love of philosophy is incompatible with feeling humiliated or trounced or arrogant or self-congratulatory for one's contributions to it – have been ignored. So this enterprise presupposes a basic and reciprocal respect for the minds, ideas and words of one's discussants, a respect that is expressed in attention to and interest in what they have to say.

Kant's concept of *Achtung* captures the intellectual attitude involved in this moral virtue of high civilization. The term is usually translated, in Kant's writings, as "respect"; and the object of *Achtung* is usually assumed to be exclusively the moral law. But Kant's account of reason in the first *Critique* makes quite clear that the moral law is not separate from the workings of theoretical reason more generally, but rather an application of it to the special case of first-personal action. On Kant's view, we feel *Achtung* toward *all* the ways in which reason regulates our activity, both mental and physical. Moreover, in the *Groundwork* Kant makes it equally clear that he is not diverging from an important common, vernacular meaning of the term, which is closer to something like "respectful attention." When you and I are trying to get clear about the implications of a statement one of us has made – when we are fully engaged in the activity of "wring[ing] philosophical insights from the English language, pressed hard and repeatedly," *Achtung* is what we feel for the intellectual process in which we are engaged and the insights we thereby bring forth.

And when Kant says that *Achtung* "impairs [*Abbruch tut*] self-love," he does not mean that *Achtung* crushes our egos or makes us feel ashamed of being the self-absorbed worms we know we are. He means, rather, that the value, significance, and power of the thing that compels our attention compels it so completely that we momentarily *forget* the constantly clamoring needs, demands and egocentric absorptions of the self; the object of our respectful attention overwhelms and silences them. For that moment we are mutually absorbed in the object of contemplation, or in actively responding to it – by acting, or by articulating it, or by evaluating its implications, or by reformulating or defending it – rather than trying to mine the discussion for transient satisfactions of our psychological cravings for self-aggrandizement. *Achtung* is an active, conative response to an abstract idea that overrides and outcompetes our subjective psychological needs as an object worthy of our attention.

These are the rare moments of intellectual self-transcendence in which together, through "extreme literal clarity, with no rhetoric and the least possible use of metaphor, with an avoidance of technical terms wherever possible, and with extreme patience in the step-by-step unfolding of the reasons that support any assertion made, together with all the qualifications that need to be added to preserve literal truth," we succeed in fashioning an idiolect subtle and flexible enough to satisfy and encompass all of the linguistic nuances we each bring to the project of verbally communicating our thoughts to each other. It is then that we achieve the only genuine unity with another of which we are capable. Alcibiades' drunken and complaining encomium to Socrates was also a eulogy to his own transient victory in achieving – even momentarily – the intellectual self-transcendence Socrates demanded.

3. Philosophical Rationality: Transpersonal or Egocentric?

Now I said that Hampshire described this Anglo-American update on the Socratic ideal as itself an historical fact. But is it? Here is a competing description of the same historical circumstance, from a rather different and less high-minded perspective:

> Victory was with those who could speak with the greatest appearance of clear, undoubting conviction and could best use the accents of infallibility. Moore ... was a great master of this method – greeting one's remarks with a gasp of incredulity – *Do* you *really* think *that*, an expression of face as if to hear such a thing said reduced him to a state of wonder verging on imbecility, with his mouth wide open and wagging his head in the negative so violently that his hair shook. "*Oh*!" he would say, goggling at you as if either you or he must be mad; and no reply was possible. Strachey's methods were different; grim silence as if such a dreadful observation was beyond comment and the less said about it the

> better .... [Woolf] was better at producing the effect that it was useless to argue with *him* than at crushing *you* .... In practice it was a kind of combat in which strength of character was really much more valuable than subtlety of mind.[5]

Here the writer is John Maynard Keynes. Where Hampshire saw the character dispositions of transpersonal rationality in full flourishing, Keynes sees psychological and emotional intimidation. Where Hampshire saw the flowering of a moral virtue of high civilization – the flowering, in Nietzsche's terms, of "slave morality," Keynes sees little more than a less-than-subtle power struggle among *Übermenschen*, driven by the instinct to win social status, even at the cost of philosophical integrity. Where Hampshire saw self-transcendence, Keynes sees egocentric rationality in full force. Who saw more clearly?

The answer is important for answering the question as to whether the character dispositions of transpersonal rationality are as central to philosophical practice as they are purported to be; and so, more generally, whether the character dispositions of transpersonal rationality *can be* as central to the structure of the self as I, in this project, argue they are. The answer to this more general question bears on the import and implications of my thesis. If philosophical practice is about the exercise of transpersonal rationality, as Hampshire suggests, and transpersonal rationality is central in the structure of the self, then philosophical practice exercises the capacity that centrally structures the self; and we cultivate and strengthen the rational dispositions of the self through philosophical practice. This confers on the philosophically inclined not special moral knowledge, but rather the special moral responsibilities of cultivating those capacities wisely and exercising them judiciously – i.e. the moral responsibilities of Plato's philosopher-king.

If, on the other hand, philosophical practice has nothing to do with transpersonal rationality and everything to do with the egocentric rationality of mutual intimidation, as Keynes seems to argue, then philosophical practice is little more than a struggle for power; and the branches of philosophy we practice are mere means to that end – no better, nobler or more indispensable than any other. Determining the type and strength of rationality in the structure of the self sheds light on the extent of our capacity for rationality in our philosophical practice, and on the legitimacy of its claim to be the "queen of the disciplines," providing method, wisdom and guidance for the process of reflection on any subject. Both of these familiar, aristocratic descriptions of philosophy convey the traditional understanding of philosophy as a noble pursuit, and impose on philosophers the moral burden of *noblesse oblige*.

---

[5]John Maynard Keynes, "My Early Beliefs," in *Two Memoirs* (New York: Augustus M. Kelley, 1949), 85 and 88; quoted in Elizabeth Anderson, *Value in Ethics and Economics* (Cambridge, Mass.: Harvard University Press, 1993), 121.

There can be little doubt that Hampshire's version of the Socratic ideal of philosophical dialogue requires of us a standard of intellectual and moral conduct to which we are, most of the time, intellectually and morally inadequate; and so that the ideal of transpersonal rationality so valorized by a "slave morality" may be – for us – little more than that. Here the moral inadequacy exacerbates the intellectual inadequacy. It is difficult enough to keep in mind at one time more than a few steps in an extended and complex philosophical argument, or fully appreciate the two opposing views that must be reconciled, or grasp the point of your opponent's criticism as he is voicing it while you are mentally both formulating your refutation of it and refining your view so as to accommodate it. But these purely intellectual limitations are made so much worse by what Kant calls "certain impulsions" of "the dear self" that obscure or interfere with the clarity and sure-footedness of the reasoning process: the need to be right or amusing at another's expense, the need to prove one's intelligence, the need to triumph, or to secure one's authority, or to prove one's superiority, or mark one's territory; or, more viciously, the need to intimidate one's opponent, to attack and crush her, shut her up, express one's contempt for her, exact revenge, teach her a lesson, or force her out of the dialogue. All of these needs exist on an ethical continuum, from the merely regrettable or pathetic at one end to the brutal or sadistic at the other. The essence of our moral inadequacy to Hampshire's Socratic ideal of philosophical conduct is our temptation to use even the limited skills of philosophical dialogue we have as a tool of self-aggrandizement or a weapon to bludgeon our opponent, rather than to arrive at recognizable truths we can both embrace.

This temptation vies with our longing for wisdom, imagination and kindness – and sometimes loses the struggle. And then it finds vivid expression in certain familiar philosophical styles most of us have encountered – or deployed – at one time or another. For example, we have all at some point surely met – or been – *the Bulldozer*. The Bulldozer talks at you, at very great length, rather than to you; and seems to understand by "philosophical dialogue" what most people understand by "lecture." Indeed, Bulldozers may make excellent lecturers, and lecturing is an excellent training ground for bulldozing. The Bulldozer expounds at length his view, its historical antecedents, and its implications; anticipates your objections to it, enumerates each one, complete with examples, and refutes them; explains the views of his opponents and critiques them; and no doubt does much, much more than this, long after you have excused yourself and backed away with a muttered apology about needing to make a phone call. Sometimes the Bulldozer seems almost to induce in himself a trance state by the sound of his own words, and seems impervious to your ineffectual attempts to get a word in edgewise. And should you momentarily succeed in getting a word in edgewise, rest assured that there will not be many of those. For any one of

them may set off a further volcanic eruption of speech in the Bulldozer, a shower of philosophical associations that must be pursued at that moment and to the fullest extent, relentlessly, wherever they may lead.

There is something alarmingly aimless and indiscriminate behind the compulsiveness of this performance, as though it were a senate filibuster without a motion on the floor; as though the Bulldozer's greatest defeat would be to cede even the tiniest corner of verbal territory to someone else. Of course the experience of "conversing with" a Bulldozer is extremely irritating and oppressive, since one is being continually stymied in one's efforts to join the issues under scrutiny and make intellectual contact with one's discussant. But I think it is not difficult for any of us to imagine how it feels to *be* a Bulldozer, to feel compelled to surround oneself stereophonically with the ongoing verbal demonstration of one's knowledge; to blanket every single square inch of the conceptual terrain, up to the horizon and beyond, with one's view of things; to fend off alien doubts, questions, and interjections of data into one's conceptual system by erecting around oneself a permanent screen of words and sounds so dense and wide that nothing and no one can penetrate it. Of course the Bulldozer himself may not think he is thwarting philosophical contact with others but instead enabling it; and may believe, even more tragically, that if he just says enough, he will surely command agreement in the end. Those many philosophers who reject the temptation to bulldoze create the necessary conditions for philosophical contact, and may even inspire *agape* – if not agreement – in their discussants.

Whereas the Bulldozer performs primarily for the sake of self-defense, *the Bully* performs more aggressively, in order to compel others' silent acquiescence; and thereby betrays her anticipation that they will speak up against her. She may deploy familiar locutions designed to forestall objections or questions before they are raised: "Surely it is obvious that ..." or "It is perfectly clear that ..." or "Well, I take it that ..." The message here is that anyone who would display such ignorance and lack of insight as to call these self-evident truths into question is too philosophically challenged to take seriously; and the intended effect is to intimidate the misguided into silence.

For example, I resorted to some of these bullying techniques earlier, in my discussion of Kant. "Kant's account of reason in the first *Critique* MAKES QUITE CLEAR that the moral law is not separate from the workings of theoretical reason more generally," I claimed; and "in the *Groundwork* Kant MAKES IT EQUALLY CLEAR that he is not diverging from an important common, vernacular meaning of the term *Achtung*." In both of these cases, I tried to double the barrage of intimidation, by brazenly combining claims of self-evidence with an appeal to authority. Why? Because even though I know these views to be controversial, I wanted you to swallow them on faith, for the moment, without questioning me, so I could go on and build on those assumptions the further points I wanted to make. Elsewhere I do argue that a

careful and unbiased look at the texts will support them. But I did not want to have to defend them here, or allow this General Introduction to the Project to turn into the exercise in Kant exegesis that I elsewhere undertake in earnest. So instead I finessed them through an attempt at intimidation; by insinuating, in effect, that ANYONE WHO'D TAKEN THE TIME TO STUDY THE TEXTS CAREFULLY could not fail to agree with my interpretation; and that any dissent from it would reveal only the dissenter's own scholarly turpitude. This is not philosophy. This is verbal abuse.

This kind of bullying may have many causes. It may result from a dispositional deficiency of self-control, i.e. of "extreme patience in step-by-step unfolding of the reasons that support any assertion made." For Hampshire does not notice that this moral virtue of high civilization may be best suited to a mild, placid, even phlegmatic temperament; and may be largely unattainable for those of us who tend toward excitability, irritability, or an impatient desire to cut to the chase. But this does not excuse the indulgence of these tendencies at your expense. After all, part of the point of philosophical training is to learn, not merely a prescribed set of texts and skills of reasoning, but also the *discipline* of philosophy. We are required to discipline our dispositions of attitude and motivation as well as of mind in its service. This is no more and no less than cultivation of the character dispositions of transpersonal rationality requires.

Philosophical bullying may also result from a negligence encouraged by the structural demands of professionalism, i.e. from a failure of intellectual discrimination. Excelling in any of the various branches of philosophy demands specialization. This may lead us to underestimate the importance of securely grounding with "step-by-step unfolding of the reasons that support" those parts of our views that lead us into other philosophical subspecialties – as, for example, political philosophy may lead into philosophy of social science, logic may lead into philosophy of language, epistemology may lead into philosophy of science, metaethics may lead into philosophical psychology, or any of these may lead into metaphysics or the history of philosophy. And since the scarcity of jobs and limited professional resources often places us in a competitive rather than a collaborative relationship with our colleagues in other subspecialties, we may be tempted, on occasion, simply to ignore, dismiss or bully our way out of the kind of careful attention to foundations that Hampshire recommends.

Furthermore, most of us entered this field because we needed to make a living doing something (true *Untertanen* that we are), and enjoyed doing philosophy enough to want to make a living doing it. As with any job on which our economic survival depends, we often have to balance the quality of our output against the time or space we have in which to produce it. We are here to ply our trade, to speak authoritatively to the designated issues. And if what we have to say depends on unfounded or insufficiently argued

assumptions, then (at least for the time being) so much the worse for those assumptions, and for those innocents who, not understanding the implicit rules of the game – the allotted speaker time, the maximum acceptable article length, or the limited market demand for fat, ponderous books such as this one – would attempt to exercise quality control by calling those assumptions into question.

The Bully becomes a morally objectionable *Überbully* with the choice of more insulting or hurtful terms of evaluation, and with the shouting, stamping of feet, or even throwing of objects that sometimes accompanies his attempts to drive home a point. This mere failure of impartiality, self-reflection and self-control shades into unadorned wrongdoing when these tactics of verbal intimidation include insinuated threats of professional retaliation or clear verbal harassment. Suggestions that holding a certain philosophical position is not conducive to tenure or reappointment, or that one will be dropped from a project for challenging received wisdom, or that raising objections to a senior colleague's view is offensive and inappropriate; as well as familiar locutions such as "Any idiot can see that ..." or, "That is the most ridiculous argument I've ever heard;" or, "What a deeply uninteresting claim;" or, "How can anyone be so dense as to believe that ...?" are all among the Überbully's arsenal of verbal ammunition. Philosophers have been publicly and professionally humiliated for having argued a view that, in their critic's eyes, marked them as dim-witted, ill-read, poorly educated, lazy, devious, evasive, superficial, dull, ridiculous, dishonest, manipulative, or any combination of the above. Whereas the Bulldozer prevents you from contributing to the dialogue, the Überbully uses you and your philosophical contributions as a punching bag, trying to knock the stuffing out of them and scatter their remains to the wind.

It is tempting to explain this grade of lethal verbal aggression as an expression of arrogance or boorishness. It is better understood as an expression of fear. Like the Bully, the Überbully attempts to demolish you through verbal harassment, not rational philosophical analysis – in clear violation of the canonical rules of philosophical discourse. All we need to ask is why either brand of bully feels the need to resort to these thuggish tactics when the canonical ones are available, in order to understand their brutal performances as an exhibition of felt philosophical inadequacy that expresses fear of professional humiliation. The frequency with which shame and fear emerge in these forms interrogates the suitability of the practice of philosophy to stand as a testimonial to our achievement of the Socratic/Hampshirean "moral virtue of high civilization," thereby as a testimonial to the victory of "slave morality," and thereby as a testimonial to the centrality of reason in the structure of the self. And it explains why my optimism about our rational capacity to transcend the merely comfortable, convenient, profitable, or gratifying is cautious at best.

The philosophical style we may describe as *the Bull* probably originates in the exhilarating discovery of esoteric knowledge that induction into any field of specialization brings. This tactic works best on students, or on colleagues who work in a different subspecialty than oneself. Like the Bulldozer and the Bullies, the Bull discourages questioning or dialogue, and silence dissent. The Bull may spew forth, with a great and rapid show of bombast, a torrent of technical or esoteric terminology, or inflated five-syllable abstractions. Or she may issue – again with no apology and much pomp – several incoherent, inconsistent, or mutually irrelevant assertions, and appear surprised at any suggestion of paradox. Or she may answer your pointed questions with a barrage of vague philosophical generalities that seem not to engage the issues at all. And the Bull may borrow some tactics from the Bully, in suggesting that any failure to grasp the overarching point of these turgid *non sequiturs* is merely a distressing symptom of your own philosophical incompetence. In this way the Bull uses the specialized tools of her trade to exclude you from participation in the private club to which she lets you know she belongs. The not-so-subtle message the Bull intends to communicate is: No Trespassers. Unlike the Bull's other philosophical utterances, this one is clear, easily grasped, and usually elicits compliance. For it is not easy to remain involved in a discussion in which the suspicion quickly grows that one's discussant is talking nonsense. Philosophers who eschew the temptations of the Bull for unvarnished clarity of exposition express the intellectual virtue of courage – the courage to expose their ideas to scrutiny without the protective pretense of intellectual superiority.

The *Bullfinch*, by contrast, simply flies away home. The Bullfinch avoids philosophical dialogue altogether, by declining to subject his own views to philosophical scrutiny or provide it to others'. Convinced of the veracity of his own views yet concerned to preserve their inviolability, the Bullfinch withdraws from philosophical engagement with unconverted others. Rather than argue his views, the Bullfinch at most will explain where he stands, ignoring retorts, criticisms or opposing views by declining to acknowledge their philosophical worth. The Bullfinch is more likely to view his own beliefs as so self-evidently true that it is beneath him to have to articulate or expose them to unconverted others in any form; and his opponent's beliefs as dangerous enough to justify getting rid of her at any cost. Thus the Bullfinch defends the sanctity of his convictions by refusing to defend them at all, instead retreating into silence, backhanded Machiavellian maneuvers, or flight. Or he may resort to cruder tools of psychological intimidation – of the sort Keynes describes – as more appropriate to his opponent. By refusing to engage in rational dialogue even as a weapon of intimidation, the Bullfinch thus approaches most nearly the explicit conduct of Nietzsche's *Übermensch*, for whom unvarnished displays of egocentric power completely replace the Socratic ideal of transpersonal rationality, and so express most clearly his

unqualified contempt for his philosophical opponents. As contempt never trumps compassion or curiosity as an intellectual virtue, the Bullfinch thereby merely confesses his felt disinclination – or inadequacy – to meet the standards of engagement that rational dialogue requires.

4. Philosophy, Power, and Historical Circumstance

These brief character sketches provide a practical counterpoint to the Socratic ideal that Hampshire describes – an ideal that finds only partial realization at best. They do not exhaust the styles and strategies of intimidating philosophical practice, and there are more lethal ones than these: to treat philosophical contributions from others as though they had not been made; or as though they had been made by someone of higher professional status; or as autobiographical rather than philosophical in import; or as symptoms of mental illness; as well as the more subtle variants Keynes describes. The common motive that underlies all of these styles of dialogue is an egocentric desire to establish and maintain hierarchical *übermenschlichen* superiority, by silencing philosophical exchange rather than inviting it. This motive is not entirely foreign to any of us. But it is meant to stifle the exercise of transpersonal rationality that seduced most working philosophers into the field to begin with, and that virtually all, with varying degrees of success, genuinely strive to practice. As such, it is, in effect, an effort to obliterate the point and practice of philosophical dialogue altogether – dialogue that indeed very often does begin with the best of intentions, reflective of the Socratic ideal which virtually all of us learned to revere as undergraduates. Philosophers who manage to persevere in the patience, generosity of spirit, and thickness of skin necessary for withstanding these assaults on the core of the practice without stooping to respond in kind are often singled out and revered for the philosophical paragon they offer to the rest of us. It is worth asking what it is about the practice or profession of philosophy in general that kindles the impulse to obliterate it; and how it is that this impulse can co-exist within the same field of inquiry as those successful practitioners of Hampshire's Socratic ideal. For this impulse does not signal merely our moral and intellectual inadequacy to the ideal. It expresses the lethal and ultimately suicidal desire to eradicate it.

We have certain external procedural devices for cloaking this suicidal impulse. There is the authoritarian device, of supplying spoken discussion with a strong-willed moderator; and the democratic device, of scrupulously invoking Robert's Rules of Order to govern every verbal contribution; and the juridical, testimony-cross-rebuttal-jury deliberation device, of the standard colloquium format. But if we were all as civilized as Hampshire's description supposes, we would not need any of these external devices. We would not need a moderator to end filibusters or umpire foul balls because no one would be tempted to hog the allotted time or hit below the belt. We would

not need Robert's Rules of Order because no one would be tempted to disrupt or exploit it. And we would not need the standard colloquium format because that format formalizes a dialectical procedure to which we would all adhere naturally and spontaneously, as do Aristotle's temperate men to the mean and Kant's perfectly rational beings to the moral law. These devices are muzzles and restraining leashes designed to rein us in, not merely from expressing our philosophical enthusiasms too vehemently or at excessive length; but rather from too obviously lunging for the jugular under the guise of philosophical critique.[6] Sometimes it is as though in our serious philosophical activity we needed to be monitored and cued from the wings by an instructor in the basics of philosophical etiquette. It is as though there were no internalized voice of intellectual conscience to guide and subdue our egocentric philosophical behavior at all.

How is this lack of philosophical self-discipline to be understood? How are we to understand the frequent identification of personal and professional wellbeing with having at least temporarily obliterated one's philosophical enemies, and of personal and professional failure with having lost the war? And how are we to understand our own self-deception and lack of insight into the egocentric motives and meaning of such philosophical behavior – as though a punishing philosophical work-over that verbally dices one's opponent into bite-size chunks were cognitively indistinguishable from the "cultivated caution and modesty in assertion" that Hampshire rightly applauds? Should we say that if we are incapable of practicing rational self-restraint and self-scrutiny in the circumscribed and rarified arena of philosophical dialogue, there is small hope for doing so in more complex fields of social interaction? Or should we say, rather, that it is because the philosophical arena is so small and morally insignificant that we have devoted so little attention to habituating ourselves to proceed in a temperate and civilized manner; and that our *übermenschlichen* barbarity here has no practical implications for our rational moral potential elsewhere?

The latter response is inadequate on several counts. First, the concept of rational philosophical dialogue as establishing metaethical conditions for comprehensive normative theory is too central to the moral and political views of too many major philosophers – Rawls, Habermas, Hare, Rorty, and Dworkin among them – to be dismissed as morally insignificant. If we cannot even succeed in discussing, in a rational and civilized manner, what we ought to do, it is not likely that we will succeed in figuring out what we ought to do, much less actually doing it. Second, talk is cheap; talk is the easy part of moral rectitude. If we can ever hold our tongue, choose our words, and exert ourselves to understand another and communicate successfully with her when our egocentric interests are at stake, then we have what it takes to

[6]So much for Hampshire's injunctions against metaphor.

cultivate the transpersonally rational character dispositions to do those things. The question then becomes whether we are less inclined to cultivate them when it is our purely philosophical interests that are at stake; and what that might reveal about the ability of philosophy – and so transpersonal rationality – to give point and form to our lives. Certainly there are those for whom philosophy is merely an intellectual game.

Third, philosophy as the transpersonally rational discipline par excellence has fashioned its own identity through the centrality of its involvement in the most elemental and universal ideals of human life – ideals of the good, the true and the beautiful; of equality, rationality and grace. These are the ideals that inspire the young to study philosophy, and that often sustain our allegiance to it as we grow older. That the intellectual skills with which we pursue research into these ideals can be so easily perverted by the Bulldozer, the Bullies, the Bull, and the Bullfinch in the service of the bad, the false and the ugly is no minor matter. How a profession self-defined by its transpersonal rationality and its idealism can generate suicidally self-repressive and self-abasing styles of professional behavior in any of its practitioners demands explanation.

Earlier I suggested that part of the explanation is to be found in the economic conditions that have come to characterize the profession of academic philosophy over the last half-century. These conditions have encouraged a possessive and authoritarian attitude toward philosophical ideas that is incompatible with the obligations of philosophical practice as Hampshire enumerates them. We have seen that these include a commitment to clarity, precision and care in the development of an argument or view; and a methodological caution that eschews easy answers for the sake of a coherent thesis that is fully cognizant of significant objections and alternatives to the view being defended. But these obligations must compete with the mounting difficulty of finding long-term or permanent jobs in the field.

Up to the early 1960s philosophy was a small, homogeneous, economically secure academic enclave. As would befit a community of *Übermenschen*, Stevenson's Emotivism vied with Ross's and Pritchard's Intuitionism and Moore's Non-Naturalism as the metaethical views of choice. Kantian, rationality-based metaethical views were not in the competition. With Johnson's Great Society programs of the mid-1960s, American philosophy began to open its doors to the ethnic, gender and class diversity among younger scholars that has always been representative of the population of the United States. But those programs in higher education funded this expanded academic population only briefly. Since then, and up through the turn of the century, the resulting scarcity of jobs has become an increasingly serious problem for younger philosophers, newcomers and legatees alike. It has been a central professional fact of life for over three decades. Those of us who entered the professional side of the field as

graduate students in the mid-1970s had studied, benefited from, and taken as role models philosophical writings that uniformly predated this dearth of professional opportunities. But we had also received a letter from the American Philosophical Association, routinely sent to all aspiring graduate students, advising them that very few jobs were likely to be available upon receipt of the Ph.D. Under these circumstances, such aspiring graduate students have had three choices: (1) ignore the letter; (2) ignore those aspects of one's previous philosophical training that conflict with it; or (3) try to adapt to both in ways that will allow one to compete successfully in the field. Clearly, the student who is both rationally self-interested and committed to philosophy will choose (3), and most who have survived professionally have done so.

For the most part the results have not been auspicious for the health of the field. The methodological caution that is essential to doing good philosophical work has been too often supplanted by an intellectual and philosophical timidity that is the antithesis of it. Understandably concerned to ensure their ability to continue and succeed professionally in the discipline to which they are committed, many younger philosophers in the past few decades have grown increasingly reluctant to fulfill the demands of the Oedipal drama that is essential to the flourishing of any intellectual discipline. In order to break new ground, younger thinkers must strive to study, absorb, elaborate, and then criticize and improve upon or replace the authoritative teachings on which their training is based. Otherwise they fail to achieve the critical independence and psychological and intellectual maturity that enable them to innovate new, stronger, and more comprehensively authoritative paradigms in their turn. Strawson's early critique of Russell's theory of descriptions, for example, or Rawls's rejection and displacement, as a young man in his early thirties, of Moore's philosophy of language-based metaethics, or Barcan Marcus' and Kripke's early repudiation of Quine's constraints on quantificational logic, or Kuhn's displacement of Popper's philosophy of science in the early 1960s are only a few of the available contemporary role models for playing out this drama in philosophy.

The obligations of philosophical practice as Epictetus and Hampshire enumerate them – and as Socrates exemplifies them – create an ideal context of transpersonal rationality within which all of the characters in this drama can thrive. In attending only to the quality of philosophical contributions and not to the hierarchical position of those who make them, the "style of philosophical talking" Hampshire describes is designed to call forth the best philosophical efforts of all parties, regardless of rank or stature. Careful, patient and rational philosophical discussion is the great equalizer among

discussants, the great leveler of professional hierarchy.[7] This is a context in which younger philosophers can feel secure in the conviction that in subjecting the views of their elders to searching scrutiny and possible refutation, they are only doing what the obligations of philosophical practice demand.

This transpersonal ideal of equality in rational dialogue comes into direct conflict with a reality in which professional survival is a scarce commodity doled out as reward in a zero-sum game among egocentrically motivated combatants. Where philosophical error translates as professional failure, the avoidance of professional failure requires the concealment of philosophical error at all costs. Under these circumstances there can be little place for the rational criticism and analysis of views, and so little place for unconstrained give-and-take among rational equals. These practices must be replaced by a system of patronage of the unempowered by the empowered, and mutual aggrandizement of the empowered by one another. It is because rational philosophical dialogue recognizes no professional hierarchy that other, extra-philosophical or even anti-philosophical measures must be invoked to maintain it under circumstances in which hierarchical status is the surest index of professional survival.

Philosophy as an academic discipline is correspondingly unusual in the obsessiveness and rigidity with which the character and composition of its traditional professional hierarchy has been guarded in recent decades. In this traditional hierarchy, with few exceptions, criticism from peers is received as an honor, whereas criticism from subordinates is resisted as insubordination; and novices, newcomers, provisional members, and interlopers tend to rank among the lowest subordinates of all. Accordingly, the more they diverge – in thought, appearance or pedigree – from the tradition, the closer to the bottom of the hierarchy they are likely to be found, and the more blatant the exercises of power that keep them there. Correspondingly more attention has been given to Kantian, rationality-based metaethical views in recent decades, and many newcomers, provisional members, and interlopers – including particularly large numbers of women – are to be found among their proponents.

Younger thinkers who choose to diverge or defect rather than conform philosophically embark on a dangerous Oedipal drama in which they must

[7]Indeed, there are few other fields in which the intellectual activity that centrally defines the discipline is so thoroughly inimical to professional hierarchy. Even in the natural sciences, such a hierarchy is justified to some extent by the training, experience and accumulation of information and methodological resources required in order to ascend to its pinnacle. Only in philosophy (and perhaps mathematics) is it possible for some unschooled pipsqueak upstart to initiate a revolution in the field with an offhand, "Here's a thought!" issued from the safe haven of the armchair. Kripke's early work in modal logic would be an example; Parfit's on personal identity would be another.

confront and face down the wrath and resistance of their elders in order to prevail. By finally rejecting the views of those whom they have studied and by whom they may have been mentored and protected in the beginning stages of their career, younger scholars will often provoke disapproval, rejection or punitive professional retaliation from those who feel betrayed by their defection. They may risk their professional survival, advancement, and the powerful professional networks that the authoritative support of their mentors has supplied. This is of course an exceedingly painful and intimidating prospect for all concerned, elders and prodigal sons[8] alike. It is nevertheless necessary in order to advance the dialogue and ensure the intellectual health of the discipline. This requires that the egocentric urge to professional self-preservation at all costs be subordinated to the demands of transpersonal rationality.

The elders will survive this defection with their stature intact – as did Russell, Moore, Quine and Popper; and eventually come to recognize their own example in that of their defectors. After all, they, too, were once defectors, and took the terrible risks of transpersonal rationality they now discourage their own disciples from taking. Thus those disciples need to demonstrate their respect for their elders, and the depth of their influence as role models, by similarly having the attachment and commitment to their own ideas, the energy and courage to probe their deepest implications, and a confidence in their value firm enough to impel them to this confrontation, despite the clear dangers to their professional self-interest. Otherwise these ideas become little more than disposable vehicles for promoting professional self-interest, of questionable value in themselves.

One might argue that this brand of naive intellectual bravado is in mercifully short supply under the best and most professionally secure of circumstances. But nerve fails all the more quickly as the threat of professional extinction becomes more real; and this failure of intellectual nerve has by now so completely pervaded the field of philosophy that it has generated its own set of professional conventions – a virtual culture of genuflection, relative to which merely to embark on the confrontation with one's elders is a serious and sometimes fatal breach of etiquette. So, to take a few examples, when I was a junior faculty member, a very senior and very eminent colleague reprimanded my efforts to defend the position developed in this project by informing me that it was "not [my] place to have views." I lost the support of a leading senior philosopher, and thereby a peer-reviewed

[8]I use this expression advisedly, since those who survive the confrontation are overwhelmingly male. The field numbers approximately 10,000 members. At last count, women occupied eight percent, and African-American women .003 percent, of all tenured positions. The punishments inflicted for their philosophical insubordination are correspondingly more virulent.

publication, by refusing to delete an example that mentioned race in a paper she had offered to recommend for publication. I once had a paper accepted for publication on the sole condition that I excise my critique of a major figure in the field; and had one rejected because a single negative referee's report, although acknowledged by the editor to be incoherent and self-contradictory, came from an important personage. Rather than take on the major thinkers, many have been encouraged or coerced by such tactics to avoid the Oedipal confrontation altogether, and diverted instead into harmless and insignificant wheel-spinning. The great, ongoing contentious debates that extended from Plato through Kant, Fichte, Hegel, Schopenhauer and on to the Vienna Circle, Russell, Wittgenstein, and Habermas seem to have been all but silenced by the repressive dictates of professionalism.

These genuflective norms of etiquette undergird the recommendations of professional self-interest, by encouraging and rewarding excessive deference to philosophical authority, by discouraging forthright argumentation and critique, and by undermining the intellectual and professional confidence of younger philosophers in their ability to develop their own views independently and survive confrontation with their elders. They thereby infantilize the powerful, by insulating their views from honest critique and thus inadvertently perpetuating the illusions of philosophical invulnerability and professional entitlement. And they infantilize the unempowered as well, by stripping them of the very resources most essential, in the long term, to their own survival and flourishing: the character dispositions of transpersonal rationality. It then would be unsurprising to discover that, when the unempowered were rewarded for their obedience with professional empowerment, the character dispositions of transpersonal rationality were given both less exercise and less philosophical weight.

These norms of genuflection, necessitated by economic imperatives, create the authoritarian conditions under which the Bulldozer, the Bullies, the Bull, and the Bullfinch can flourish. Like other artifacts of the culture of genuflection, they function to protect canonical or insecure philosophical territory using anti-philosophical weaponry, when pure philosophical dialogue itself is too subversive of established hierarchy or received interpretation to be tolerated. And through practice, repetition, and professional reward, these repressive philosophical styles are transmitted as role models from one generation of graduate students to the next, as legitimate modes of philosophical discourse. Ultimately they supplant the legitimate and civilized modes of philosophical discourse Hampshire describes with self-aggrandizing displays of power and domination, and corrupt the quality of philosophical ideas accordingly. In replacing the transpersonal obligations of philosophical practice with the egocentric imperatives of professional survival, these styles bespeak more than our self-centeredness. They bespeak our inability to transcend structural conflicts

between the democratic prerequisites of a genuine philosophical meritocracy and the inequitable consequences of a market economy.

5. Philosophy as Exemplar of Transpersonal Rationality

Western philosophy has always found its source of value in its identification with transpersonal rationality, originally the systematic rational inquiry practiced by Socrates. But as other disciplines – the natural sciences, psychology, sociology, political theory, anthropology – have gradually seceded from the formal discipline of philosophy and formulated their own rational methodologies, philosophy has repeatedly sought outside itself for its defining exemplar of rationality, and so for its source of intrinsic value. Up through the nineteenth century, Anglo-American analytic philosophy ignored the defection of the natural and social sciences and identified rationality with empirical rational inquiry, i.e. with scientific methodology. Traditional epistemology began to be upstaged by the newly emerging subspecialty of philosophy of science. At the beginning of the twentieth century, the melding of logic and mathematics in Russell and Whitehead's *Principia Mathematica* provided philosophy with another exemplar of transpersonal rationality with which to identify: one of logical rigor, symbol and system. Traditional speculative metaphysics received a corresponding boost in status at the same time that it took a drubbing from Logical Positivism. After the Second World War, philosophy turned to Frege, Wittgenstein and Chomsky for yet another exemplar of rational philosophical method as linguistic analysis. Linguistic anthropology and sociology received correspondingly more attention from philosophers of language. And over the last two decades of the twentieth century, philosophy increasingly turned back to the sciences – this time to the emerging field of cognitive science – for its exemplar of rational methodology. The philosophy of mind and theory of action have flourished accordingly. Trade relations have thus run in both directions: the discipline of philosophy has exported and diversified its early conception of transpersonal rationality as systematic Socratic inquiry into newly emerging research disciplines; and these, in turn, import back into the discipline of philosophy more highly specialized conceptions of their own.

The more the discipline of philosophy has succumbed to the political, economic, and professional pressures just described, the more stridently it has insisted upon these externally imported exemplars – sometimes singly, sometimes in tandem – as centrally definitive of the field and the practice of philosophy. And the more the discipline of philosophy as the practice of transpersonal rationality *par excellence* has been threatened from any and all directions, and the more the specialized conceptions of rational methodology have proliferated, the more tenaciously philosophy has held onto its self-identification with transpersonal rationality as such, adjusting its source of value according to how in particular transpersonal rationality is conceived.

In the end, however, it is only philosophy's original identification with the systematic rational inquiry of Socrates – Epictetus' injunction to

> transgress in nothing the clear pronouncement of reason ... to live as one who is mature and proficient, and let all that seems best to you be a law that you cannot transgress. ... [to] attend to nothing but reason in all that [you] encounte[r]. ... to live as one who would wish to be a Socrates[9]

that remains impervious to defection, attack, or nonrational alternatives. It is impervious to defection because emerging fields that have defected have taken rational Socratic inquiry with them as their minimal foundations. It is impervious to attack because any such attack must presuppose its methods in order to be rationally intelligible. And it is impervious to nonrational alternatives because no such alternative competes with it on its own ground. Philosophy's greatest challenge, then, is to live up to its traditional, Socratic self-conception: conduct in all spheres that accords centrality to the character dispositions of transpersonal rationality.

Under the historical circumstances earlier described, it is impossible to avoid calling into question the present-day adequacy of philosophy to meet this challenge, and so its right to insist on its self-definition as an exemplar of transpersonal rationality. Hence it is impossible to avoid questioning whether the character dispositions of transpersonal rationality can be as central to the structure of the self as they seemed to have been for Socrates and Epictetus. The problem would seem to be not that we so often violate Epictetus' injunction to "transgress in nothing the clear pronouncement of reason;" but rather that we so often transgress that clear pronouncement in precisely those areas of conduct in which reason is purported to reign supreme. One explanation would be Keynesian: that philosophers have been guilty of self-serving pretensions to rationality all along; and that philosophical practice has never consisted in anything more than psychological intimidation and the flouting of power imbalances under the guise of rational dialogue. According to this view, Epictetus' entreaties would be addressed precisely to those in need of transpersonal rationality as an inspiring ideal by which to moderate largely egocentric behavior.

But another possibility is that we must rather take special care now, at the turn of the twenty-first century, to defend the centrality to philosophy of those character dispositions of transpersonal rationality the exercise of which have been so traditionally definitive of its practice. It might be that these dispositions, and so the traditional practice of philosophy itself – and so its adequacy as an exemplar of transpersonal rationality – are now under particularly severe attack, from both inside and outside the discipline, by concerted attempts to defend traditional power relations against the radically destabilizing effects of rational Socratic interrogation. The displacement of

---

[9]*Op. cit.* Footnote 1.

transpersonal rationality from a central functional and valuational role in the way the structure of the self is conceived signals a move away from the "slave morality" that valorizes the character dispositions of transpersonal rationality as essentially constitutive of human survival and flourishing. This displacement also signals a move toward alternative, *übermenschlichen* norms of egocentric behavior that implicitly condone freer and more blatant exercises of power in the service of desire, instinct and emotion. It is no accident that this Gestalt shift occurs at an historical juncture when such exercises and displays of power are increasingly necessary to defend conventional social arrangements – both inside and outside the academy – against rational Socratic interrogation by individuals and communities traditionally disempowered by them; and are valorized by unconstrained market forces that dismantle the democratic underpinnings of the social contract. But it is then doubly ironical that the character dispositions of transpersonal rationality themselves should be marshaled by some philosophers to justify them.

The philosophical use of reason to justify unreason then obliges those philosophers who explicitly value reason, rational interrogation, and the character dispositions of transpersonal rationality more generally as intrinsic goods to defend them in turn. It requires us to reaffirm and protect these intrinsic goods as essential and definitive of philosophical practice, regardless of the express philosophical views on which they are honed. It requires us as well to realize these values in our philosophical practice, regardless of professional repercussions. And it requires us to disregard those repercussions as secondary to the preservation of rational integrity. That is, the philosophical task is to demonstrate the deeply entrenched necessity of transpersonal rationality to coherent thought and action, independently of the express metaethical views or valuation of rationality any particular philosopher might hold. That is my task in this project.

6. The Enterprise of Socratic Metaethics

In ethics we distinguish between a normative and a metaethical theory. A *normative* moral theory tells us what we ought to do, and why. Thus it traditionally utilizes such prescriptive terms as "ought," "should, "good," "right," "valuable," or "desirable." I offer an analysis of such terms in Volume II. This is the *practical* part of a normative theory, also known as *casuistry*. Such a theory also contains a *value-theoretic* component that enlists certain states, conditions, or events that explain what *is* good, right, or desirable: friendship, for example; or love, or reason, or integrity. Value theories differ with respect to both content and structure; I say more about these distinctions in Chapter V of Volume I.

By contrast, a *metaethical* theory seeks to unpack the metaphysical presuppositions of a normative theory: to what sorts of entities, if any, its

prescriptive terms refer; whether it can be objectively true or not; what its scope of application might be; what conception of the agent, rationality, or human psychology it presupposes. Thus a metaethical theory is descriptive and analytical where a normative one is prescriptive and hortatory.

By comparison with the putative centrality of transpersonal rationality to the practice of philosophy itself, the metaethical views philosophers expressly defend show a much wider range of variation in the role each assigns to rationality in the structure of the self. Here the value and function of reason ranges from the central to the peripheral, and the prominence of nonrational elements in the view's conception of the self varies accordingly. At one extreme, consider *Subjectivism*. Subjectivism is a radically Anti-Rationalist view that essentially rejects truth and objectivity as possible goals for intellectual discourse on any subject. But any judgment in the categorical indicative mood implies – whether rightly or wrongly – the truth and objectivity of the judgment, including the judgment that truth and objectivity are impossible. So if that judgment, that truth and objectivity are impossible, is itself true and objectively valid, then it is false and objectively invalid. If it is false, then its negation, i.e. that truth and objectivity are not impossible, is true. So the truth of Subjectivism implies its falsity. If, on the other hand, Subjectivism is neither true nor false, then it refers to nothing and expresses at best the speaker's emotional despair about the possibility of communication – a condition treated better in psychotherapy than in intellectual discourse. If this paradox of judgment strikes you as in any way troubling, or as detracting from the intelligibility of Subjectivism, then you have already accepted intellectual criteria of rational consistency that imply an aspiration to objective validity and truth. Only when these criteria are presupposed can meaningful or coherent discussion, on any topic whatsoever, proceed.

A fortiori, any judgment of specifically moral value aspires to be more than a mere emotive expression of the speaker's momentary feelings. It aspires to objective validity, and we signal this by stating our views publicly, defending them with evidence or reasoning, and subjecting them to critical analysis in light of standards of rationality and truth we implicitly accept. So, for example, suppose someone walks up to you and punches you in the nose. Your verbal reaction will surely include the statements that he had no right to do that, that his behavior was unwarranted and inappropriate, and that you did nothing to deserve it. It is not likely that you will then go on to add that of course these are just your opinions which have no objective validity and that there is no final truth of the matter. Rather, you express your beliefs in categorical indicative judgments, which you of course presume to be true, and which you can defend by appeal to facts you take to be obvious and values you take to be equally obvious. Of course some of your presumptive judgments may be mistaken or false. But this does not entail that there is no fact of the matter as to whether they are or not.

The project of moral communication has not only to do with letting others know what we think, but also trying to command their acknowledgement that we are right. Those of us committed to the Socratic ideal prefer to command this acknowledgment through rational dialogue rather than emotional rhetoric, dissimulation, psychological manipulation, or threats of professional or social rewards withheld or punishments inflicted for dissenting. That is, we do our best to "live as one who would wish to be Socrates," rather than as a Bulldozer, Bully, Überbully, Bull, or Bullfinch. By relying on the force of rational dialogue to win agreement with our moral convictions, we try to command not only others' assent, but also their intellectual respect. In rational discussion, analysis and argument, we reach beyond the circle of the converted to try and convert the unconvinced. We express respect for the transpersonally rational capacity of the unconverted by appealing to it, rather than to their emotional, psychological or social vulnerabilities, to convince them. And we receive the best confirmation of the truth of our moral convictions when others are rationally convinced, rather than manipulated or coerced or deceived, into adopting them. Call this the enterprise of *Socratic metaethics*. Socratic metaethics grounds moral convictions and judgments in the Socratic ideal of rational dialogue as a means for arriving at moral truth.

Within the enterprise of Socratic metaethics, there are many ways to proceed. One that has a long historical pedigree is what I shall call Humean Anti-Rationalism, because it takes its inspiration from the authoritative status Hume assigns to desire and the passions in justifying moral action.[10] In earlier historical periods this approach emerged variously in normative theories such as Intuitionism or the Moral Sentiment Theory of the British Moralists. (Similarly, Virtue Theory claims allegiance to Aristotle, but on extremely shaky exegetical grounds). As developed in the early twentieth century philosophy of Sir David Ross, Intuitionism stipulates the existence of an innate faculty of moral intuition, consultation of which tells us what moral principles we ought to follow in action.[11] Prominent late twentieth century Humean Anti-Rationalists such as Annette Baier, Lawrence Blum, Michael Stocker, or Susan Wolf harken back to British Moralists such as Shaftesbury, Hutcheson, or directly to the Hume of Book III of the *Treatise*, by repudiating the governing role of moral principle and instead appealing to moral emotion or sentiment to guide action.[12] Similarly, the Noncognitivism of Allan

[10]This is Thomas Nagel's term to characterize variants on the same group of views I discuss here. See his *The Possibility of Altruism* (Oxford: Oxford University Press, 1975), 8. I devote Chapter VII in Volume I to study of this work.

[11]Sir David Ross, *The Right and the Good* (Oxford: Clarendon Press, 1938).

[12]Annette Baier, *Moral Prejudices* (Cambridge: Harvard University Press, 1994); Lawrence Blum, *Friendship, Altruism and Morality* (Boston: Routledge and Kegan Paul, 1980); Michael Stocker, *Valuing Emotions* (New York: Cambridge University Press,

Gibbard, Joseph Raz, and Elizabeth Anderson rejects the rationality of moral principle – but then resurrects rationality as a prescriptive criterion for moral emotions and attitudes. In all of these cases, moral guidance is given by a nonrational component of the self: We ought to perform those actions we intuitively know to be right, or, respectively, feel most deeply. No consistent Humean Anti-Rationalist normative view can have a developed practical or casuistical component, because what any particular individual ought to do depends on their particular intuitions, feelings, or desires – not on impartially conceived principles. Nevertheless, the value-theoretic parts of these views are articulated and developed within the impartial normative constraints of Socratic metaethics.

Volume I will contain much, and Volume II a slight bit more, on the failings of late twentieth century Humean Anti-Rationalism. Here I call attention to just one reason why it is unpalatable *in practice* to anyone seriously interested in the enterprise of Socratic metaethics as a distinctive philosophical methodology. This is that it appeals to the authority of a first-personal, interpersonally inaccessible experience in judging, not only what *one* should do, but what should be done *simpliciter* under particular circumstances. In consulting only one's moral emotions or intuitions about how to resolve some hypothetical or actual moral problem that need bear no obvious or articulable relation to one's own circumstances, one presumes to legislate how others should behave or feel on the basis of a moral foundation which is cognitively inaccessible to them, and therefore inaccessible to their evaluation.

Suppose, for example, that I discover that my best friend is dealing drugs to minors and decide, on the basis of my feelings about him, to protect our friendship rather than betray it by turning him in to the police. There is a great deal you and I may discuss about such a case. But without knowing, and without being able to experience directly the particular nature and quality of my feelings for this person, you may find my behavior simply indefensible. You may acknowledge and sympathize with the deep bonds of friendship and loyalty I am feeling, but find it nevertheless impossible to condone my claim that I just could not bring myself to destroy them by turning him in. You may think that no friendship, no matter how deep or meaningful, should count for so much that it outweighs the right of minors to be shielded from drug addiction before they are mature enough to make a rational choice. And since I cannot convey to you the direct quality of the

---

1996); Susan Wolf, "Moral Saints," *The Journal of Philosophy 79*, 8 (1982); First Earl of Shaftesbury, "Selections," in *The British Moralists: 1650 – 1800* (Oxford: Clarendon Press, 1969); Francis Hutcheson, *Illustrations of the Moral Sense*, Ed. Bernard Peach (Cambridge, Mass.: Belknap Press of Harvard University, 1971); Hume, *A Treatise of Human Nature*, Ed. L. A. Selby-Bigge (Oxford: Clarendon Press, 1978), Book III.

experience of my friend on which my feelings are based, there is little I can say to defend my decision. Perhaps I may expect your pity or sympathy for my dilemma, but I cannot expect your respect or agreement. So unless you find me particularly compelling as a role model on nonrational grounds (say, my crucial presence in your upbringing; or my charisma, or broad sphere of social or professional influence; or your desire to stay in my good graces), I can provide you with no reason why the principles on which I acted (and even Humean Anti-Rationalists act on principles, even if they don't think about or formulate them) should govern your behavior under similar circumstances.

This is not a peculiarly Kantian objection. Unless a principle on which I act is formulated partially, i.e. with indexical operators, proper names or definite descriptions, we presume it to apply impartially; that is the way language works. Terms and principles have general application to the scope of referents they denote, unless their scopes are restricted explicitly by stipulation or fiat or context. So, for example, if I tell you that dogs are susceptible to gastric tortion, I am either mistaken or else using the term "dog" in an idiosyncratically restricted sense, to refer specifically to large dogs with cylindrical stomachs. Similarly, if I tell you I feel that friendship should come before social welfare, you will naturally take me to be doing more than merely emoting my personal feelings about this particular friend. You will naturally take me to be expressing a judgment that applies not only to my own behavior in this case, but to anyone's who must weigh the relative priority of friendship and social welfare. But since I am merely telling you what I feel, and since what I feel is not directly available to you, I offer you no available justificatory basis for evaluating the applicability of this principle to your behavior. Unless you have some special reason to be impressed with my feelings, you have no reason to be impressed with the principles on which I act. Late twentieth century Humean Anti-Rationalism, then, subverts in practice the enterprise of Socratic metaethics on which it relies in theory, by appealing to interpersonally inaccessible moral states to justify its moral judgments.

Ross's Intuitionism was couched in a metaethics that attempted to avoid this outcome, and more recent Humean Anti-Rationalists may adopt a similar strategy. Ross argued that the principles we morally intuit as the outcome of careful and considered reflection on the circumstances in question were objectively valid, in the same way that mathematical intuitionists argue that the objects of mathematical intuition, such as the basic truths of arithmetic, are objectively valid. But this makes intuition, as well as its objects, even more cryptic and cognitively inaccessible than before: What if we have different moral intuitions about the same case? What if yours puts social welfare ahead of friendship? How do we determine which one of us is morally defective, and in what respect? The difficulty Intuitionists face in claiming an

objectively valid status for the moral judgments they make is that intersubjective agreement can provide the only evidence for the mysterious mental capacities required to make them; and this, of course, makes the enterprise of Socratic metaethics itself unnecessary. Where rational dialogue becomes necessary to addressing the unconverted that lie outside one's circle of sympathizers, Intuitionism has nothing to say.

Some late twentieth century Humean Anti-Rationalists have adopted a similar strategy, by claiming a certain veracity for moral emotions, based on their authenticity as a forthright expression of a person's most centrally defining values and projects. This resolves Humean Anti-Rationalism into a species of Subjectivism: If a certain judgment authentically expresses my centrally defining values and projects, it is true, at least for me. I do not think this is an interesting use of the term "true," and will not pause to rehearse any more of the elementary objections to Subjectivism. Suffice it to raise the obvious problem, analogous to that faced by the Intuitionist, of how to dispose of the authentic feelings and judgments of the unconverted; or of a storm trooper or lynch mob. Otherwise the basic objection stands: late twentieth century Humean Anti-Rationalism appeals for its persuasive power on interpersonally inaccessible moral states, and thereby sabotages the enterprise of Socratic metaethics on which it relies.

By contrast, *Rationalism* takes the enterprise of Socratic metaethics seriously as a methodological presupposition of *all* metaethics. The method of Rationalism is to try to justify a moral theory or principle by appeal to reason and argument as the currency of interpersonal communication. A Rationalist seeks to lead her reader or listener from weak and mutually acceptable premises to a substantive conclusion as to the most convincing substantive moral theory or principle, by way of argument, analysis, critique, and example interpersonally accessible to both. A Rationalist may appeal to imagination, personal experience, or certain feelings or perceptions or intuitions as reasons for or against a particular view; but she views reason – not the feelings or perceptions or intuitions or other responses invoked *as* reasons – as the final arbiter of rational dialogue.

In this undertaking, Rationalism is neither broadly democratic nor narrowly fascistic. A Rationalist does not try to gain adherents for her view by oversimplifying the theory or the arguments, or by obfuscating them with neologisms or inflated prose or verbal abuse or grim silence in order to intimidate others into accepting it. In appealing to reason, Rationalism addresses itself only to those who are willing to exercise theirs. It does, however, assume that all competent adults can do so, *regardless of culture or environment*. In this it is more democratic than Humean Anti-Rationalism, which demands intersubjective concurrence in substantive moral judgment as the only convincing evidence of the truth of those judgments, when in fact there is no necessary connection between intersubjective concurrence and

truth at all. For these among other reasons, Rationalism defines the critical methodology adopted in this project. The argument proceeds by appeal to reasons and critical analysis, and most of the philosophers discussed here proceed similarly in defending their views – regardless of the substantive content of those views.

7. Rationality and the Structure of the Self

The main focus of discussion in this project is with two competing branches of Rationalism, prevalent in mid- to late twentieth century Anglo-American analytic philosophy, that differ with respect to the role each assigns to rationality in the structure of the self. Both branches agree upon the Socratic metaethical enterprise as a philosophical methodology. Both agree, as well, on the necessity of providing a metaethical conception of the subject as agent, as a foundation for making normative claims about what subjects as agents should do. And both agree upon the necessity of explaining what they think moves subjects as agents to act, and in what they think acting rationally consists. But each branch deploys different models of human motivation and rationality as the shared, weak metaethical premises on the basis of which to argue for these normative moral claims. The first branch is what I call the *Humean conception of the self*, the second the *Kantian*. Thus both Humean and Kantian conceptions *in fact* count as varieties of Rationalism according to this taxonomy, regardless of the Anti-Rationalist content some Humean views may have.

7.1. Two Conceptions of the Self

By a *conception of the self*, I mean an explanatory theoretical model of the self that describes its dynamics and structure. A conception of the self is to be distinguished from a *self-conception*, which is the same as a "personal self-image." The latter expresses the way or ways in which an individual thinks of himself, for example, as nice, well-intentioned, grumpy, loyal, fastidious, etc. It typically plays a normative role in individual psychology: We try to live up to the ideal individual we conceive ourselves to be, and regard negative attributes as flaws or deviations from that ideal. Thus a self-conception is part of one's normative moral theory. By contrast, a conception of the self plays a descriptive, metaethical role in moral theory: It identifies and describes the kind of individual to whom the theory purports to apply. For example, a normative moral theory that urges general conformity to the Golden Rule on the metaethical grounds that it best enables each individual to promote her self-interest implicitly identifies those individuals to whom the theory is addressed as desiring to promote their self-interest. Similarly, a normative moral theory that recommends actions governed by the dictates of reason metaethically presupposes reason as a significant motivational factor in the relevant agents.

Traditionally, moral philosophers who write systematically and discursively always begin by describing their conception of human subjects as agents before they tell us what they think those agents ought to do. That is, they preface their normative claims with a metaethical conception of the self to which those claims are intended to apply. If they did not, we would have no way of gauging whether or not we ourselves were intended subjects of the theory. A conception of the self, then, provides a metaethical account of the psychological facts about human agents considered as subjects of normative moral principles.

My question in this project is not that of which normative moral theory is uniquely correct. It is the more foundational question of which metaethical conception of the self underlying normative moral theories provides the most accurate account of the psychological facts. If a moral theory's underlying conception of the self is fallacious or largely inaccurate regarding the psychology of human nature, the question of the theory's validity for human beings can scarcely arise.

A conception of the self as I define it comprises two parts: First, it includes a *motivational model*. This explains what causes the self to act, and how. It identifies those events and states within the subject that constitute its capacity for agency; and it explains how, under certain specified conditions, those capacities are realized in agency. So the motivational model in a conception of the self is an explanatory and causal model. The motivational model with which we are most familiar and comfortable is the Humean, belief-desire model of motivation, according to which we perform those actions we believe best satisfy the desires that move us.

Second, a conception of the self includes a *structural model*. This describes and charts the conditions of rational coherence and equilibrium within the self. It depicts that state of the self in which it functions as a unified psychological entity, and maintains psychological balance and integrity among its cognitive and conative components. Again the structural model we largely take for granted is the Humean, utility-maximizing model of rationality, according to which all of our actions aim to maximize satisfaction of our desires; I described this earlier as egocentric rationality. Taken together, the structural and the motivational models of a conception of the self explain what a unified subject is and how it is transformed into responsible agency.

The Humean and the Kantian conceptions of the self are each grounded to some extent, although not entirely, in the writings of Hume and Kant respectively. The first has been the prevailing conception within Anglo-American analytic philosophy at least since Sidgwick: Humean premises concerning motivation and rationality are now widely accepted in such disparate fields as psychology, economics, decision theory, political theory, sociology, and, of course, philosophy. The Humean conception is engendered

by, but is not identical to, Hume's own conception of the self. Nor is it embraced in its entirety by any one of its adherents. Rather, different facets of it are pressed into service to do different philosophical jobs: to explain behavior, for example; or predict preferences; or to analyze moral motivation, or freedom of the will. Thus the picture I sketch in Volume I is a composite one, drawn from many different sources in mid- to late twentieth century philosophy. This conception has been refined and elaborated to a high degree of detail in decision theory and the philosophy of mind, and its theoretical simplicity and apparent explanatory potency is attractive. These are serious and impressive achievements with which any sustained critique of the Humean conception must directly engage. But it has resulted in simplistic approaches to the understanding of human behavior in the social sciences, and it has generated enormous problems for moral philosophy. – This, shortly put, is the critical view I defend in Volume I. I offer arguments that systematically unpack some of the major internal and functional defects of the prevailing Humean conception of the self, with an eye to later highlighting the superior comprehensiveness, explanatory force, and suitability for moral theory of its proposed rival.

The second branch of Rationalism in moral philosophy is less popular: Kantian premises regarding motivation and rationality are accepted in some areas of moral philosophy, social theory, and cognitive psychology, but are not widely shared outside them. I believe that the full power of this conception of the self has not been sufficiently explored or exploited, and in Volume II I try to begin to remedy this. Relative to the enterprise of Socratic metaethics, my fundamental – but not my only – objection to the Humean conception of the self, and consequent allegiance to the Kantian, can be summarized quite simply: By insisting on desire as the sole cause of human action, the Humean conception of the self limits our capacity for action to the comfortable, convenient, profitable, or gratifying; and correspondingly limits our rational capacities to the instrumental roles of facilitating and rationalizing those egocentric pursuits. The Humean conception thereby diminishes our conception of ourselves as rational agents, by failing to recognize or respect the ability of transpersonally rational analysis and dialogue, as described above, to causally influence our behavior, even as it deploys and depends on them in philosophical discourse. This immediately raises the question, unanswerable within the traditional framework of metaethics itself, of what Humean moral philosophers take themselves to be accomplishing by discursively and rationally elaborating their views in print. If transpersonal rationality is incapable of changing minds or motivating action, as Humeans frequently claim, what is the point of deploying it to defend their views in books, articles and symposia? Or is the point merely to get tenure and attract disciples motivated similarly by careerist considerations to adopt and promulgate those views? Whereas Humean Anti-

Rationalism subverts the enterprise of Socratic metaethics in practice while relying on it in theory, the Humean conception of the self subverts Socratic metaethics in theory while relying on it in practice. If the Humean conception of the self is right, then the practice of philosophy is little more than an *übermenschliches* power game. But if that conception is wrong or incomplete, then Humeans are ignoring the larger arena in which these little games are played out.

7.2. Volume I: The Humean Conception

Essentially, Volume I of this project complains about other people's views, including, of course, Hume's own. It nevertheless expresses *Achtung* for these views, and for the thought and hard work that went into them, by treating each in depth rather than in passing. Its critical arguments are intended to motivate us to rethink our commitment to the prevailing Humean paradigm, first by pointing out defects in its twentieth century formulation and use in metaethical justification; and second, by scrutinizing the extent to which we may validly appeal to the authority of history and tradition in support of that formulation. I try on the one hand to acknowledge the technical sophistication and practical power of the Humean conception, and on the other to call attention to certain formal and theoretical limitations that I believe require the detailed treatment that I try to give them. I suggest that this conception is in fact a special case of an alternative, transpersonal conception of the role of reason – the Kantian conception that I elaborate in detail in Volume II – that is broader in scope, more firmly ensconced in the traditional canon, and more radical in its implications for practice.

7.2.1. The Two Models

Taken together, the belief-desire model of motivation and the utility-maximizing model of rationality constitute the Humean conception of the self as driven by desire to maximize the satisfaction of desire under all circumstances. I begin by considering separately each of the two models that comprise the Humean conception: first the belief-desire model of motivation in Chapter II, then the utility-maximizing model of rationality in Chapters III and IV. Here my focus is on the internal, structural defects of these models themselves, irrespective of their deployment in any particular moral theory. I base my formulation of the belief-desire model on the classic discussions of Brandt and Kim, Goldman, and Lewis; revise and refine it in light of certain problems that arise within that classical formulation; and elaborate some of the further problems, both structural and metaethical, that even that sympathetic reformulation cannot avoid. In Chapters III and IV I give the same detailed attention to the utility-maximizing model of rationality, and argue in Chapter IV that even the sophisticated mid-century reformulations and formal elaborations of this model undertaken by Von Neumann-

Morgenstern, Allais, Ramsey, Savage, and others do not avoid its intrinsic structural defects. I conclude that the structural defects of the Humean conception of the self more generally can be avoided only by resituating it as a special case within the more comprehensive, Kantian conception of the self discussed in Volume II.

By scrutinizing the problems and flaws inherent in the Humean conception itself, Chapters II through IV prepare the ground for the criticisms in Chapters V through XIV, of some of the myriad ways in which this conception of the self has been pressed into service to provide formally sophisticated and scientifically reliable foundations for a wide variety of twentieth century normative moral theories. I begin this survey in Chapter V, by dislodging my subsequent examination of these theories from the straitjacket into which Anscombe's influential distinction between consequentialist and deontological theories has forced them. I argue that this distinction obscures rather than illuminates the complex structure of a fully developed normative theory; and that so-called consequentialist moral theories are in fact merely Humean exemplars in disguise. I reject Anscombe's obfuscating distinction in order to focus more sharply, in the rest of Volume I, on the actual, detailed structure and content of some of those leading late twentieth century moral theories that – regardless of their stated allegiance – depend on Humean metaethics, without the benefit of Kantian presuppositions. All, whether they identify themselves as Humeans, Kantians, New Kantians, Anti-Rationalists or Noncognitivists, make use of the Humean models of motivation and rationality as foundational justificatory premises for their normative moral theories. I argue that all such theories founder on the inadequacy of these models to the task.

7.2.2. Three Metaethical Problems

Late twentieth century normative moral theories that invoke the Humean conception of the self as a justificatory foundation thereby engender three fundamental metaethical problems that each one of these theories then tries to solve, and that are insoluble within its own confines:

(1) First there is the problem of *moral motivation*: Can moral considerations alone move us to act in others' interests? The belief-desire model of motivation implies that they cannot; for that model stipulates that all action is motivated by the pursuit of desire-satisfaction, and only desires have causal influence on action. This means that rational appeals, argument and dialogue by themselves are *in theory* insufficient to reform, change minds, create desires, or inspire action. Hence on the Humean conception of the self, specifically philosophical dialogue alone is equally impotent to reform the culpable. Chapter VI defends this conclusion, as well as this formulation of the problem of moral motivation, against Humeans who declare that there is no such problem because the belief-desire model of motivation is compatible

with moral motivation as that term is ordinarily understood. Chapter VII then examines in depth Thomas Nagel's classic effort to substantiate this declaration by grafting a Kantian account of moral motivation onto a Humean foundation. Nagel's is not the only attempt to demonstrate the compatibility of this odd couple; but it was the first, the most thorough and the most original. All later efforts take their cue from Nagel's resourceful analysis. I argue that it fails to reconcile them, but succeeds in laying the groundwork for an alternative, truly Kantian solution to the problem of moral motivation.

(2) The problem of *rational final ends* is connected with (1): Can reason identify any alternative final ends independent of desire-satisfaction – for example, altruistic or transpersonal moral ones, that it would be rational for us to adopt? According to the utility-maximizing model of rationality, it cannot; only desire can play this role, and reason has a merely instrumental function. Hence philosophical reasoning is incapable of articulating viable alternative visions of the good – of virtuous character, for example, or of a good life – that diverge from those we have been conditioned or hard-wired to accept. Chapter VIII defends this conclusion by criticizing four interconnected, prominent late twentieth century Humean and Anti-Rationalist attempts to solve the problem of rational final ends within the constraints of the Humean conception. I argue that neither Frankfurt nor Watson offer viable solutions to the infinite regress of higher-order desires that threatens a Humean account of self-evaluation. And neither Williams nor Slote offer convincing accounts of personally inviolable ground projects, in the absence of transpersonally rational criteria for identifying and evaluating those final ends. However, all four call attention to important dimensions of personal ethics that an adequate solution to the problem of rational final ends must accommodate.

(3) The problem of *moral justification* is, in turn, a special case of (2): In propounding a particular moral theory using the familiar philosophical tools of discursive reasoning, moral philosophers undertake to demonstrate the transpersonal rationality of a particular end or value or vision of the good, i.e. that value-theoretic set of social arrangements or principles of action prescribed by their theory. Moral justification stands at the intersection between normative ethics and metaethics. For just as a theory's practical part tells us what we ought to do and its value-theoretic part explains why so doing is worthwhile, similarly its moral justification is meant to rationally convince us to adopt the values that confer worth on the actions thus prescribed. It thus appeals to metaethical considerations of transpersonal rationality that may require us to transcend the valued arrangements and ends with which we already may be comfortable, in order grasp the value of others which may be unfamiliar. But if reason itself can neither motivate us to adopt the valued arrangements prescribed by such a theory as an alternative

final end, nor justify our doing so, then either these arrangements must be justified instrumentally, as in some sense a means to desire-satisfaction; or else they cannot be rationally justified at all – in which case the enterprise of substantive moral philosophy, and the acknowledged standards of transpersonal rationality that guide it, are futile.

Chapter IX criticizes three Humean varieties of metaethical justification that wrestle with this dilemma: Noncognitivism, Deductivism, and Instrumentalism. I argue that Anderson's Noncognitivist theory of value reduces to a conformist and socially conservative, Rawlsian conception of interpersonal validation; that Gewirth's ambitious and comprehensive Deductivist justification of his Principle of Generic Consistency is subverted by his allegiance to the belief-desire model of motivation; and that the utility-maximizing strategy of Instrumentalist justification deployed by Rawls, Brandt, Gauthier, Harsanyi and others is inherently self-defeating. Chapters X and XI then examine two of the most prominent Instrumentalists – Rawls and Brandt – in depth. I show, first, that the Humean structural similarities between their attempts at justification override their contrasting ideological allegiances; second, that both founder on exactly the same Humean vulnerabilities; and third, that both thereby illuminate some of the pitfalls that a satisfactory solution to the problem of moral justification must avoid.

Chapter XII then applies these conclusions to the most quintessentially Humean normative moral theory. Classical Utilitarianism presupposes the belief-desire model of motivation in its conception of human agency, and the utility-maximizing model of rationality in its Instrumentalist metaethical justification. This theory received its most rigorous formulation from Sidgwick at the turn of the twentieth century, and its most significant mid- to late century refinements from Hodgson, Gibbard, and Lewis. But the insolubility of the Free Rider problem within these constraints demonstrates that Humean Instrumentalism is no more conceptually coherent at the level of normative moral theory than it is at the level of metaethical justification. I argue that each one of the above normative moral theories contains much to recommend it. But all of them come to grief over their Humean assumptions about justification.

Thus I conclude that the above three problems – of moral motivation, rational final ends, and moral justification – can be solved only by replacing the unreconstructed Humean conception with a more comprehensive, Kantian conception of the self which the Humean conception, suitably reconstructed, implicitly presupposes. So my approach to refuting Humeans is in the end the same as Kant's to refuting Hume: essentially to accept much of what Hume said, but then to articulate the necessary foundational presuppositions that enabled him to say it.

7.2.3. Hume Himself

Attempts are often made to counter the above objections to the Humean conception of the self by appeal to Hume's own authority. In particular, it is sometimes suggested that, despite superficial textual appearances to the contrary, Hume's model of rationality does *not* imply that rational action consists simply in satisfying one's desires as efficiently as possible, whatever they may be; and hence that the Humean model does not have the further counterintuitive consequence of identifying as rational actions that show a clear degree of irresponsibility or psychological instability. Rather, it is maintained that Hume did supply an account of rational final ends in his discussion of the calm passions and "steady and general view" that corrects the biases and contingencies of an individual's desires and perceptions; and that contemporary Humeans often implicitly presuppose this account. If true, this would mean that it was consistent with the Humean conception to impose special motivational restrictions on rational choosers in order to justify a moral theory, so long as these were compatible with such a steady and general view; hence that the above objections to the motivational and structural models of the Humean conception were directed against a straw man. Volume I therefore concludes with an examination of the original source of the Humean conception, and considers whether close attention to Hume's own writings – whether by his most able proponent or by me – deflects the above criticisms. Chapter XIII examines Annette Baier's thoroughgoing defense and exegetical revision of Hume. I show that, just as Kant incorporated Hume's insights into a yet broader and more subtle conception of the self, Baier's own defense of Hume similarly presupposes the very Kantian conception of the self she purports to reject. Chapter XIV then argues that a direct and detailed reconstruction of Hume's own views on these matters that considers *all* the relevant passages does not support the claim that he supplied an account of rational final ends. Instead, they undermine it. Hence the counterintuitive implications of Hume's own metaethics remain, as do the above objections to its use in justifying a normative moral theory. Finally Chapter XV summarizes and tracks the interconnections among the many Humean dogmas that have shaped the landscape of late twentieth century Anglo-American analytic philosophy, and thereby sets the stage for their refutation in Volume II.

7.3. Volume II: A Kantian Conception

Volume II contends that after having devoted two and a half centuries of attention to the Humean conception, it is now time to move on to a sustained consideration of the historically more recent, philosophically more sophisticated conception of the self that Kant proposed in response to these problems (which he, unlike we, saw right away). This conception offers a solution to the above three problems that incorporates the prevailing

Humean conception as a special case, but supercedes it as an independent explanatory and prescriptive model. The proposed Kantian conception consists not in two separate models, one of motivation and one of egocentric rationality; but rather of a single model, of transpersonal rationality, that has both motivational and structural functions in the self. This model comprises the familiar, canonical principles of theoretical reason that govern the dispositions of transpersonal rationality. So at least on the face of it, this alternative conception of the self is prettier, simpler, weaker, and more comprehensive than the Humean conception. I try to show that it is also more predictively powerful, more formally sophisticated, more entrenched canonically, and truer to the empirical facts about human agents.

Relative to the indubitable achievements of the Humean lineage in the twentieth century, a Kantian may seem to be at a disadvantage in this pursuit. Because Kant himself was out of favor in Anglo-American analytic philosophy until well after the Second World War, there is no longstanding canonical tradition, comparable to that of the Humean Utilitarian tradition in contemporary moral philosophy, of an extensively developed terminology or set of highly refined concepts, principles, formalizations, or theoretical structures on which Kantians can rely for a background frame of reference relative to which the analysis is situated. Some have raised serious questions about those that have been proposed.[13] However, this absence of a developed canonical framework is proving to be tremendously fertile and stimulating for the groundbreaking work in moral philosophy that already has brought Kant's views into the context of contemporary philosophical debate. Under the tutelage of John Rawls's lectures on Kant, [14] many of his students and advocates have ably and amply demonstrated the potential of Kant's program for contemporary moral philosophy. I join this glacial process of collaborative refinement and elaboration of the Kantian alternative that has already begun, not only in moral philosophy but also in certain branches of cognitive psychology and social theory as well.

---

[13] Elijah Millgram, "Does the Categorical Imperative Give Rise to a Contradiction in the Will?" *The Philosophical Review 112*, 4 (October 2003), 525 – 560.

[14] Edited by Barbara Herman and reprinted in Rawls, *Lectures on the History of Moral Philosophy* (Cambridge, Mass.: Harvard University Press, 2000). As it happens, my main contact with Rawls' reading of Kant was in the abbreviated form in which he presented it in his Social and Political Philosophy course, which I first took and then taught as a teaching assistant. My own Kantian educational influences – Phillip Zohn, Michael Levin, Arthur Collins, Dieter Henrich – all focused on scholarly exegesis of the *Critique of Pure Reason*. This may account for the difference in my approach to Kant in the context of contemporary moral philosophy.

7.3.1. A First Critique *Analysis of Transpersonal Rationality*

My approach in the second volume of this project differs from those of other contemporary Kantian moral views, in several respects. First, as indicated above, I reject the thoroughgoing distinction between theoretical and practical reason that other such views take for granted. Second, therefore, I do not assume that a proposed Kantian conception of the self might be developed upon the foundations of Kant's moral writings alone. Rather, I believe that Kant intended these subsequent writings to presuppose the fully articulated conceptions of the self and rationality he first developed in depth in *The Critique of Pure Reason*. Third, therefore, like Kant's own conception of the self, my contemporary refinement of it gives priority to the canons of classical logic as providing the underlying structure by which the psychological coherence and conative power of the self and intellect can be evaluated. I try to clarify some of the potentials and limitations of the Kantian conception of transpersonal rationality – for example, its capacity for establishing cognitive and psychological coherence on the one hand, and for fostering self-deception, particularly about moral action, on the other.

Thus the discussion is divided into two Parts – Ideals and Realities – in order first to elaborate in detail what the unimpeded functioning of such a self would look like; and then to use that ideal as a criterion of performance against which the malfunctions of actual selves can be explained as deviations. Just as Chapter V of Volume I had to dislodge the Humean conception of the self from the death-grip of the consequentialist-deontological distinction in ethics in order to take a fresh look at its metaethical function in twentieth century moral philosophy, Chapter II of Volume II similarly must begin by rescuing the proposed Kantian conception of the self from the clutches of the inferentialist-representationalist debate in the philosophy of language. This clears the way for a defense of the thesis that transpersonal principles of theoretical rationality are much more deeply embedded in the structure of the self than the Humean conception acknowledges; and that satisfaction of these principles is a necessary condition of psychological integrity, consistent experience, and unified agency. I propose two constraints that encapsulate these requirements: *horizontal* and *vertical consistency*; and certain modifications in classical predicate logic notation needed in order to symbolize them subsententially. Chapter III applies these modifications to rational choice notation, and thereby generates a *variable term calculus* that formally exposes the intensionality and logical inconsistency of a cyclical preference ordering; defines a genuinely rational preference; and so shows how standard decision theory, and the Humean utility-maximizing model of rationality more generally, can be fully integrated into this more comprehensive Kantian model as a special case. Chapter IV provides a test case for this conclusion in examination of a contemporary, self-described Humean decision theory.

Contrasting my approach to rational choice with Edward McClennen's, I argue that his analysis of resolute choice in fact does not depend on the Humean conception to which he professes allegiance. On the contrary, it expresses a deeper, basically Kantian conception of transpersonal rationality.

Chapter V then addresses the problem of moral motivation, and shows how the transpersonal principles of rationality developed in Chapters II and III directly cause action without any necessary intervention of desire; how they function descriptively as explanatory and predictive principles for a fully rational agent of the sort described by Kant's normative moral theory; and finally contrasts the psychology of an agent motivated by egocentric rationality with that of an agent motivated by transpersonal rationality. Chapter VI then applies this account of transpersonal motivation to an analysis of the moral emotion of compassion, and argues that far from excluding impartiality, as Humean Anti-Rationalists such as Lawrence Blum claim, true compassion presupposes it.

7.3.2. A First Critique *Analysis of Pseudorationality*

Part II of Volume II addresses the ways in which we systematically deviate from the ideal of transpersonal rationality described in Part I. Here, too, Kant's account of the synthetic unity of apperception in the first *Critique's* Transcendental Deduction is the inspiration. For if a necessary condition of unified selfhood is its internal horizontal and vertical consistency, then the self is disposed to preserve that consistency – i.e. is disposed to literal self-preservation – against anything that threatens it. And then anomalous data that defies conceptualization in terms of our familiar categories of thought truly must be for us "nothing but a blind play of representations, that is, less even than a dream," as Kant claims at A 112. In that case the gap between what we actually perceive, feel and do on the one hand, and how we conceive of those events on the other is bridged only when those events can be made horizontally and vertically consistent with our conceptions, and not otherwise.

In Chapters VII and VIII I focus particularly on the case – basically Aristotle's intemperate character – in which the motivational efficacy of the intellect is overridden by stronger forces, and the agent's will intellectually reconfigured to accommodate them, producing pseudorational apologia and ideologies that excuse these deviations from rationality to self, to conscience and to others. The concept of *pseudorationality* introduced in Chapter VII refers to the ways in which we systematically and ruthlessly force those events into the Procrustean bed of our preconceptions, ignoring or butchering or distorting them to fit the requirements of literal self-preservation. Chapter VIII applies this analysis of pseudorationality to the case of greatest interest for moral theory: that in which the anomalous events in question are our own, first-personal desires, emotions and actions. Chapter VIII also offers a

solution to the problem of rational final ends that subjects all such ends to the transpersonal requirements of horizontal and vertical consistency, and rejects as irrational those which violate them. By thus tempering and qualifying the account of moral motivation proposed in Chapter V, these two chapters serve as the foundation for the analyses in the chapters to come, of how we wrestle with the practical applications of normative moral theory.

Chapter IX addresses the problem of moral justification, by showing that Kant's analyses of commands, imperatives, and the moral "ought" reveals the psychologically and morally ambivalent relationship we bear to normative moral theory; and hence that moral justification is equivalent to causal explanation only so long as we have reason to preserve the self-conception a moral theory such as Kant's enshrines. To the extent that we do not, the project of moral justification itself becomes both more urgent and more futile. Chapters X and XI then extend this analysis of pseudorationality to the third-personal case, in which the moral anomaly – hence the threat to literal self-preservation – is not oneself but rather another. Chapter X considers the problem of moral interpretation, i.e. how the demands of literal self-preservation may combine with the tendency to pseudorationality to distort and constrict the scope of one's favored moral theory and thus produce xenophobic and politically discriminatory moral judgments of another's behavior; and suggests some further practical criteria any such theory must meet in order to restore its proper scope of inclusiveness. Although the analysis here does not furnish a metaethical justification for any one particular moral theory, it does imply that only a Kantian-type moral theory satisfies all of these criteria. Finally, Chapter XI presses our pathological motives for thus distorting the scope of our normative moral theories to their foundation, in considerations of literal self-preservation and the threats that theoretically anomalous agents represent to it; and suggests some ways in which we might restore moral inclusiveness consistently with protecting rational intelligibility.

7.3.3. Some Advantages and Limitations of the Kantian Alternative

In a nutshell, the formal difference between the Kantian conception of the self I defend in Volume II and the Humean conception criticized in Volume I is that the latter, having overlooked the traditional strengths and resources of classical predicate logic, reduces to tautology when it reaches for universality. The former, by contrast, exploits those strengths and resources to propose a way in which the latter, when properly contextualized, might partake of the nonvacuous universalization to which it aspires. The Kantian conception is thus both an alternative to and also more comprehensive than the prevailing Humean one, because it both recognizes and incorporates the data the Humean conception excludes, and also preserves its aspiration to rational intelligibility, i.e. to explanatory theoretical completeness, despite this. It

shows, first, how transpersonal rationality can be motivationally effective in action, hence that the belief-desire model of motivation is incomplete; second, that transpersonal rationality does imply substantive constraints on final ends that differentiate rational from irrational ones, hence that the utility-maximizing model of rationality is incomplete; and third, that transpersonal rationality can therefore justify a certain range of moral theories as rational final ends, and can motivate us to adopt them.

Fourth, however, reason cannot demonstrate any one of these moral theories to be uniquely rational, nor to be implied by the requirements of transpersonal rationality itself. Rather, the appeal to reason, on which we as philosophers implicitly rely, presupposes a view of ourselves as socialized moral agents who are transpersonally rational and therefore morally responsible. This view, in turn, finally presupposes a Kantian conception of the self as motivated and structured by the requirements of transpersonal rationality, to which each of the moral theories within this range implicitly subscribes.

This conception of the self opposes not only the Humean dictum that transpersonal rationality is impotent to determine the ends we seek. It also opposes the Humean Anti-Rationalist stance that treats transpersonal rationality in action as an impediment to personal authenticity. I give particular attention to whistleblowers, from Socrates forward to the contemporary context, who have marshaled the reserves of transpersonal rationality to transcend the egocentric pursuits of self-interest, the gratification of desire, and the expression of instinct and emotion, in the service of an inclusive understanding of the good in the realization of which all can cooperate. It is here that Kant joins Hobbes in rejecting Nietzsche's *Übermensch*. A social order (however well serviced by *Untertanen* blinded by "slave morality") in which all fully empowered citizens were free to wield power in the service of their instincts and desires would be no viable social order at all.

These substantive arguments are intended to present an alternative way of conceptualizing our own behavior and conscious life as better suited not only to our aims in moral philosophy, but to explanation of the psychological facts as well. The claim is, then, that our *de facto* commitment to this view of moral agency, plus the descriptive Kantian conception of the self that encapsulates it, jointly explain our actual behavior, including our reflective philosophical behavior, better than the prevailing, unreconstructed Humean alternative; and therefore provides a more realistic and appropriate justificatory foundation for moral theory.

For of course Humean moral philosophers have other reasons for rationally defending their views in books and articles besides getting tenure and attracting disciples. Like Kantians, and like most philosophers, they appeal to rational argument to convince us because they believe in the

rationality of their views. Rational considerations can cause a change not only of mind or heart. They also can cause a change in behavior as well. They can change what we teach, what we say, how we comport ourselves, and – at the very least, for whom we vote. A Kantian conception of the self acknowledges the motivational influence of rational argument on action from the outset. In speech and writing, Kantian moral philosophers exploit rationality unapologetically, through appeals to conscience and reason, and reminders of who and what we are and where our responsibilities as rational agents lie. The challenge Kantian metaethics faces is then to articulate convincingly the metaethical conception of the self, rationality, and motivation that best explains its practical import. Volume II attempts to meet this challenge.

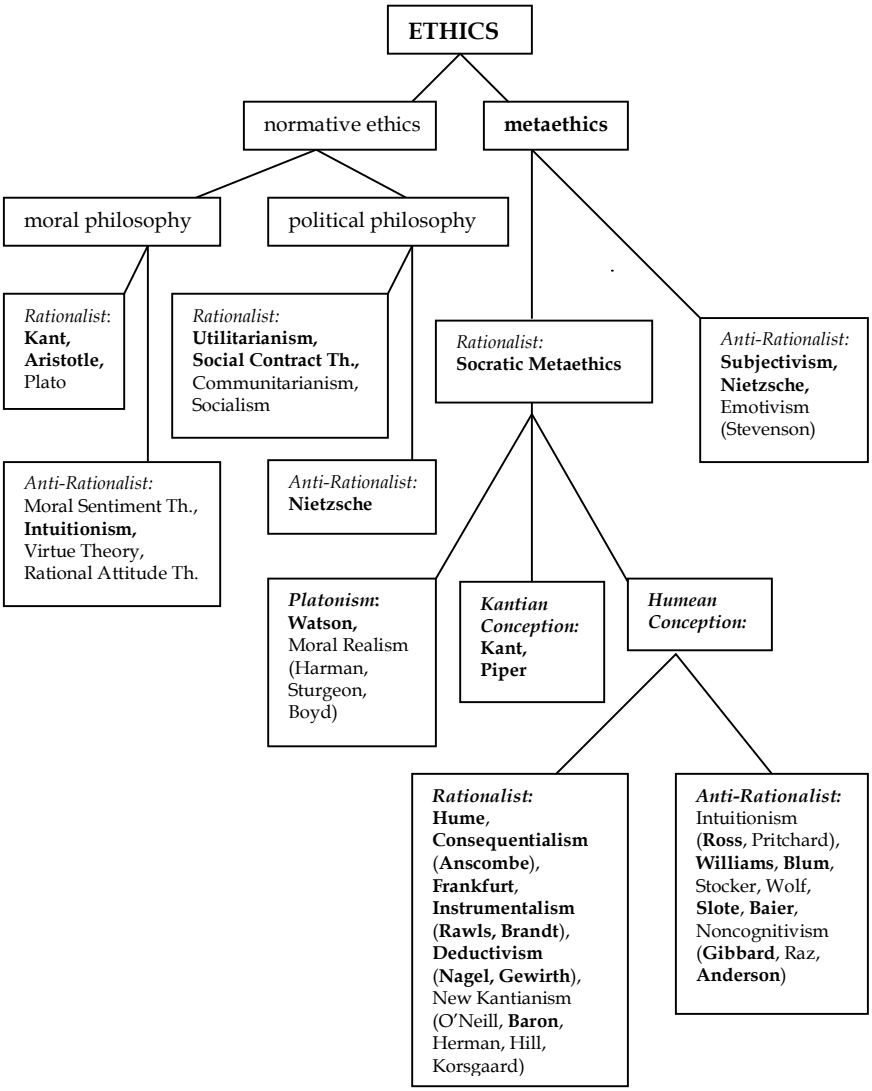
I do not expect that any of these lines of argument will necessarily compel all, or perhaps even most, Humeans and Humean Anti-Rationalists to see the error of their ways or reform them accordingly. For in the end these arguments presuppose the *value* of transpersonal rationality as the defining element in the structure and conation of the self. They presuppose that one is prepared, not only to recognize transpersonal rationality as definitive, but also to valorize its character dispositions, as a "slave morality" does. As in any philosophical disagreement, philosophical opponents may ascribe to the same rational consideration very different weights, and what is a conclusive reason to one may be an irrelevant *non sequitur* to another:

THE KANTIAN:	THE ANTI-RATIONALIST:	THE HUMEAN:
But X is *irrational*!	But X is *irrational*!	But X is *irrational*!
But Y is *counterintuitive*!	**But Y is *counterintuitive*!**	But Y is *counterintuitive!*
But Z is *unsatisfying*!	But Z is *unsatisfying*!	**But Z is *unsatisfying*!**

So even if I succeed in making a plausible case that reason has this centrality in the structure of the self, I have still relied on and presupposed the value of the very capacity I mean in my argument to valorize. A *real* Humean Anti-Rationalist who disparages the value of transpersonal rationality will therefore accord little value to my transpersonally rational arguments that transpersonal rationality has value. Indeed, I will have trouble getting her to read this project. If my reader is a real Humean Rationalist, for whom transpersonal rationality has value but no motivational efficacy, my arguments will then provide him no motivation to rethink his values, no matter how persuasive those arguments may be. Perhaps only Hobbes' astute – and rationally persuasive – observations on the necessary transience and

instability of accumulated power might lead him to reconsider the value of the Socratic ideal.

One final caveat. Volume II covers a great deal of territory. Some readers may experience it as a free fall off a steep cliff; a plunge from the metaethical paradise of philosophy of language, logic, and decision theory with which I begin into the casuistical netherworld of xenophobia and political discrimination with which I conclude. I try to maximize the reader's attention to the connections and continuities between these extremes, so as to minimize the bumpiness of the ride down. But such readers are advised to fasten their seatbelts nevertheless.



Views discussed in this project are in boldface.

Figure 2. A Taxonomy of Ethics

Chapter II. The Belief-Desire Model of Motivation

Both the motivational and the structural models, and so the Humean conception more generally, assume the concept of a desire as a value-neutral foundation, on the basis of which an equally value-neutral explanatory paradigm can be constructed. The belief-desire model of motivation is deployed in a variety of late twentieth century, foundational projects in action theory and philosophy of mind; but it is not well-defined. This chapter begins by taking Humean desire theorists at their word, with respect both to their analysis of desire and to their varying and sometimes conflicting accounts of how it functions in motivation and in explanation. Section 1.1 takes Brandt and Kim's pioneering account to task for achieving vacuity rather than the universality at which it aims. I show that regardless of whether the concept of desire is construed experientially or theoretically, the belief-desire model of motivation conceived universalistically is either false or vacuous. Section 1.2 discusses Alvin Goldman's *A Theory of Human Action*, which invokes this model in explicating the nomological connection between intentional behavior and its causes. However, I contend that because this model identifies actions in terms of the desires they are intended to satisfy, it provides us with no resources for identifying the motivationally effective desire independently of the action it is presumed to cause. Therefore it cannot explain action because it fails to preserve the distinction between *explanandum* and *explanans*. Section 1.3 then examines David Lewis' argument in "Radical Interpretation," which invokes the belief-desire model of motivation in order to explain how we can understand an agents' beliefs, desires, and meanings as he would express them in his language and as we would express them in ours, given the physical facts about that agent. But because Lewis assumes the universality of the model at the outset, i.e. that we interpret the agent as maximizing utility in his physical behavior in the same way and under the same conditions that we would in ours, no independent understanding of that agent's beliefs, desires, and meanings, as he would express them in his language, can be achieved. In all three of these cases, I argue, the problem is the same: Either a desire is just one motivational variable among many other possible ones, in which case it cannot provide a comprehensive explanatory model; or else it is a fully comprehensive and tautological "theoretical construct," in which case it fails to explain any action independently at all.

Sections 2.1 and 2.2 offer a representational theory of desire that attempts to redress these problems, within the constraints of the Humean conception. I show that this one avoids tautology and retains explanatory potency by forsaking the claim to universality, and more accurately characterizes the conception of desire that underpins the Humean conception of the self. On this alternative analysis, it is the exclusion from the self of the object of desire that gives the desirings of the self their motivational and structural

importance for the Humean conception. We are moved to action, on this view, and to regard our internal and external resources instrumentally, not because of what we are and have, but because of what we conceive ourselves literally to want, i.e. lack.

However, in Sections 2.3 through 3 I show that this more detailed account of the structure and dynamics of desire shapes a necessarily theory-laden and morally egocentric psychology. The moral psychology of desire casts human agents as driven by dissatisfaction, insecurity, and deep feelings of inferiority in all of their actions; as permanently trapped in unrealizable fantasies of future fulfillment; and as incapable of such basic human cognitive achievements as impersonality, impartiality, or self-reflection. Section 4 considers briefly the likelihood that the resulting profile will be familiar and depressing in several respects. Those who recognize themselves in it may be tempted to conclude to its practical accuracy, and therefore to the veracity of the Humean model of motivation. Those who do not, or who find it deficient or incomplete in its portrayal of the reality and potential of human motivation may rightly conclude that the belief-desire model of motivation is inadequate to the psychological facts, makes false predictions about human behavior, and hence is badly in need of repair. So there is some reason to doubt whether the Humean conception can be, not only adequate to the complexity of human behavior, but whether, indeed, it can function as an adequate explanatory paradigm at all.

A terminological point (with a familiar Kantian ring): To describe the self as having desires may seem to suggest that desires relate to the self as properties of it. But this does not follow, since of course any such nominative can occupy the place of predicate or subject indifferently. In Volume II I defend the familiar first *Critique* view that locutions such as "I desire x," or "I have a desire for x" express a relation of the unexplicated concept of the self or "I" to the experience of desiring, or of having a desire that is essentially possessive: Any such experience – of desiring, having, believing, and so on – itself has the feature of being had by someone. That is, it is true by definition of the cover term "experience" that each experience belongs to some self. I defer further amplification of these points to Volume II, Chapter II. For now notice just that the possessive relation of the concept of the self to its experiences such as desirings does not imply that the relation of actual selves to their experiences is similarly and inevitably possessive.

The Humean conception of the self implies, rather, that the relation of an actual self to its motivationally effective desires is one of identity; and its relation to its other experiences – the bundle of impressions, ideas and inferences of which Hume himself spoke – instrumental. The dispositional or occurrent experience of desiring defines the Humean self, in that it determines (1) the *structural* relations among these other internal components of the self, as nested instrumental resources for the satisfaction of desire arranged

through similarly instrumental reasoning and calculation; and (2) the *motivation* for any intentional behavior in which the Humean self engages. To suggest that the Humean self is to be identified with its desires rather than its impressions and ideas is thus to take seriously the intrinsic connections between selfhood and agency, and between selfhood and determinate psychological structure. The Humean conception of the self illustrates one way in which the question as to the structural and motivational origins of behavior can be answered.

1. Orthodox and Revisionist Variants

1.1. Brandt and Kim's Ambivalence

In characterizing what a *desire* is, Humeans converge in adopting definitions that are something like what Brandt and Kim seem to mean by a "want"; i.e. a disposition to feel pleasure or satisfaction in thinking about or admiring the object of desire, and a disposition to feel disappointment or frustration in its nonattainment.[1] Brandt and Kim explicitly mean to construe wants or desires as theoretical constructs, with no experiential analogues.[2] This interpretation allows them to apply the concept of a want or desire to the explanation of a broader range of behavior than would be suggested by the ordinary sense. It attempts to divest the concept of a want or desire of the particular experience (or conjunction of experiences) that individuate it from other motivational states. However, five of their six proposed criteria for the correct usage of "*x* wants *p*" make explicit references to *x*'s experience of such feelings as joy or disappointment in the attainment or nonattainment of *p*, pleasure in entertaining the thought of *p* or in the occurrence of *p*, and an impulse to do the act that *x* believes will eventuate in *p*. To analyze the concept of a want or desire for *p* in terms of joy or pleasure at the satisfaction of that want and a felt impulse to achieve that satisfaction seems inconsistent with denying that "want" denotes an experience. If it denotes a constellation of experiences then presumably its denotation includes each conjointly in that constellation. This ambiguity reflects a deeper ambivalence among Humeans such as Thomas Nagel, Alan Gewirth and Brandt about just what kind of entity a desire is supposed to be, and what kind of function in a conception of the self it is supposed to have.

Brandt and Kim's definition is thus of interest for its systematic attempt to capture the dichotomy expressed in Nagel's distinction between

[1]Brandt and Kim, "Wants as Explanations of Action," in N. S. Care and C. Landesman, Eds. *Readings in the Theory of Action* (Bloomington, Ind.: Indiana University Press, 1969), 199-213. References to this article are parenthesized in the text. Notice that this definition differs in several respects from that which Brandt offers in *A Theory of the Good and the Right* (Oxford: Clarendon Press, 1979).

[2]See Care and Landesman, *ibid*., pp. 200-202 and footnote 2.

unmotivated and motivated desires (discussed in Chapter VI.2.3, below), i.e. to encapsulate both the commonsensical notion of desire as an empirical event, and also those modifications of it that attempt to accommodate the demands of a theoretical first principle.[3] So the Humean conception of the self vacillates between two interpretations of what a desire is. On one view, desires are events in the world with causal power. Call this the *orthodox* view. The orthodox view admits of competing beliefs about the content of a desire, of the sort that might separate the Freudian from the Adlerian. It accepts the possibility of disagreement about what desire one has; but broaches none about the presupposition that we are in fact moved by real events called "desires." In this respect, desires are as ontologically basic as furniture or plane crashes: they occur. Although they intend otherwise, Brandt and Kim's account of desire in fact conforms to the orthodox interpretation.

The other view withholds the ascription of ontological irreducibility, in order to extend the scope of application of the term "desire" to cover the large variety of motives – e.g. greed, compassion, self-interest, duty – we commonly assume to exist. Call this the *revisionist* view. The idea of the revisionist view is that belief-desire talk supplies us with a kind of "theoretical construct" in terms of which all intentional behavior can be retrospectively described. Thus we ascribe to an agent a desire to achieve the end her behavior seems to us most clearly designed to achieve, and the belief that the action she actually performed was the most efficient way at her disposal for achieving it. These ascriptions enable us to preserve the assumption of the agent's instrumental rationality, by treating desires as ubiquitous in the sense that Nagel's motivated desires are: Whatever the agent does is assumed to promote the satisfaction of some desire she has as efficiently as she can, given the information and resources at her disposal. Brandt and Kim's account aspires to conform to the revisionist interpretation, although in fact it does not.

In the following chapter I elaborate at greater length on some of the problems with this view of desires as theoretically ubiquitous. Here merely note the resulting tension between the orthodox and the revisionist views. The credibility of the claim that desires are ubiquitous diminishes in inverse proportion to the extent to which we are required to construe "desire" as denoting an actual event or state of affairs in the world. For to that extent, it seems self-evident that on any nonvacuous interpretation of the word "desire," desires are neither necessary nor precipitating causes of many of the actions we perform.[4] In the course of this project I examine several cases in which one performs actions intentionally, not because one desires their ends,

---

[3] *Ibid.*

[4] W. D. Falk also makes this point in "'Ought' and Motivation," *Proceedings of the Aristotelian Society*, New Series, NLVIII (1954-58). See especially pp. 115-117.

but rather because they instantiate normative principles or values to which one is deeply committed.

1.2. Goldman's Orthodoxy

Some orthodox desire-theorists claim ubiquity for desire despite the plethora of counterexamples. Alvin Goldman, for example, concurs with Brandt, Kim, Gewirth, Davidson, and others in characterizing an occurrent desire or want as a kind of "pro-attitude:"

> [W]ants need not be intense or emotion-laden; they need not absorb one's whole consciousness... wanting x is roughly equivalent with [*sic*] 'feeling favorably toward x,' 'being inclined toward x,' 'being pro x,' 'finding x an attractive possibility,' 'finding x to be a "fitting" or "appropriate" possibility," etc. ... wants have various strengths or intensities.[5]

On Goldman's account, wants are mental states that, other things equal, are partial causes of basic act-tokens. Although we at present have neither techniques adequate for measuring the intensity of various wants, nor precise universal laws for predicting the acts they cause, we can often state ex post facto the beliefs and desires operative in causing behavior (73). Here the ceteris paribus clause is crucial: if other things are not equal, then the want in question may not cause the act in question: if, for example, a stronger want overrides it, or if a sudden, fleeting want momentarily eradicates awareness of it; or if external circumstances prevent one's acting on it; or if one's beliefs about how best to satisfy the want do not include this particular act (108, 113-114). But even in these cases it is, on Goldman's analysis – as it is on Nagel's, "a logical truth that certain wants (together with certain beliefs) will lead to certain behavior." Even if one particular want cannot always be counted on to cause a particular action, that action must, by definition, "have been caused by some want or other" (115).

But which want? How do we identify the motivationally effective want independently of the action itself? Goldman offers five empirical criteria of identification for the particular want that motivated an action that are, as he says, "extremely helpful in narrowing down precisely what the agent's goals or purposes are" (117):

> (1) *co-temporal acts*, such as the agent's looking in a certain direction (for example, looking at the lamp while flipping the switch as evidence that he wanted to turn on the light);
>
> (2) *sequential acts*, all of which are part of a program of action aimed at the same goal (as when, for example, the agent's looking at the lamp,

[5]Alvin Goldman, *A Theory of Human Action* (New Jersey: Prentice-Hall, 1970), 49-50. Henceforth all references to this work are parenthesized in the text.

getting up and going over to the light switch, and finally flipping the switch are all presumably aimed at turning on the light);

(3) *observable events, characteristics, or clues* given by the agent, such as facial expression, that enable us to tell what the agent's wants are;

(4) *information about the agent's likes, dislikes, and personality traits*, obtained by observation of previous behavior; and

(5) *information about the agent's beliefs*, obtained by observation of his perceiving something to be the case, and/or by his avowals.

However, first, criteria (1), (2), (4), and (5) presuppose that prior criteria for identifying the goal of an action have already been satisfied. Once we have identified the agent as, for example, looking in a certain direction rather than rotating his neck muscles ((1)); or flipping the switch rather than patting the wall ((2)); or having certain likes ((4)) or beliefs ((5)) on the basis of the actions he performs that express them, we can then invoke this information in order to identify further ones. But if the mystery is how to identify the goal of any such action itself, these criteria are unlikely to be of assistance. Goldman's empirical criteria of want-identification do not necessarily hold in all cases.

The second, more crucial difficulty is that they fail to identify the agent's wants independently of the action taken to satisfy them. Even if they enable us to identify the goal of a particular action the agent performed, and so the action itself (e.g. turning on the light rather than exploring the objects in the environment), they do not enable us to surmise what want caused him to perform it (to read legibly? to check the electrical wiring?). Of course if the motivationally effective want is stipulated to be just the desire to perform the action one in fact performed, then identifying the action thereby identifies the want, and no further question remains to be asked. But Goldman has already rejected this stipulation, and rightly so, on the grounds that the goal that identifies the action is not necessarily the goal one wanted to achieve by performing it (115). The major difficulty with these criteria, then, is that they do not enable us to preserve both ubiquity and full-blooded causal efficacy in any significant sense simultaneously. Desires cannot be both ubiquitous and ontologically basic.

So the same conclusion holds here for Goldman that I later show to hold for Nagel and Gewirth as well: To the extent that desires are events, as the orthodox view requires, they cannot be required as a necessary condition of action. In particular, desire is an unnecessary presupposition of conforming to operative social norms of the sort discussed in greater detail in Volume II.[6] For this is to interpose a superfluous intentional state – a desire, "pro-

[6]Goldman acknowledges that his analysis may not work for habitual behavior, but neglects to say how habitual and nonhabitual behavior are to be distinguished: by token? by type? by structure? by repetition? etc.

attitude," or "appetition" – between us and the behavior we perceive as normatively appropriate. More generally, to stipulate any such desire or interest that necessarily intervenes between the awareness of what a situation requires and the resulting action is both counterintuitive from the point of view of commonsense introspection, and methodologically messy. For the stipulation of such a desire as necessary in all cases is based on the self-fulfilling hypothesis that there must have been a desire present in order for one to act at all. So if my performing the action makes it true by definition that I desired its end, yet I find no evidence of such a desire when I examine carefully my own motives, then the concept of a necessarily motivating desire must be relegated to the explanatory status of a "theoretical construct." But this is to *abandon* the orthodox for the revisionist view of desire.

Now some may find it difficult, if not impossible, to examine their own motives without coming upon a desire somewhere in the mix, and it is worth asking why, and for whom this ontological ubiquity actually obtains. For example, among the operative social norms which govern our behavior are to be found, first, prevailing linguistic practices which, since the seventeenth century, have relied increasingly on a psychological vocabulary of individual desires and interests that in the twentieth century was extended even beyond the scope of conscious awareness. Second, these linguistic practices are mutually interdependent with a globally disseminated economic explanation of human motivation that gives prime emphasis to the pursuit of individual gratification, rewards, and advantages. Third, these practices and explanations are the expression of an optimistic and future-oriented system of political and social institutions that accords central recognition to individual freedom and autonomy – specifically, the freedom and autonomy to pursue the gratification of individual desires. The belief-desire model of motivation, and the Humean conception of the self more generally, has had a central place in the intellectual history of Western culture of the last three centuries.[7] So it is hardly surprising that we buy into it. We conceive ourselves individualistically, define ourselves in terms of personal desires, and esteem ourselves to the extent that we satisfy them.

Among the Tabwa of West Africa, on the other hand, observers' queries as to what an agent "wants" or "desires" are typically answered by elaborate descriptions of what normally obtains under such circumstances. Pressing the question meets with incomprehension. The Tabwa do not seem to have our concept of a desire at all.[8] Here the problem is not that the revisionist view has been mistaken for the orthodox view. It is that the concept of desire, *even on*

[7]See Albert Hirschman, *The Passions and the Interests* (Princeton: Princeton University Press, 1977).

[8]Here I am indebted to Kit Roberts for making available to me some results of her field research.

the orthodox view, is unintelligible to the Tabwa. There are no ontologically basic events in their experience that answer to our orthodox conception of a desire. Perhaps we could eventually get a Tabwan to "recognize" the existence of such events, by schooling him in our linguistic conventions, and convincing him to accept the resulting reconceptualization of his experience as valid, in the way that B. F. Farrell has documented for the indoctrination of psychoanalytic concepts.[9] But in this case we would merely have won him over to our side. We would not have demonstrated that our side has any greater claim to objective accuracy than his from an unbiased perspective. The Tabwa constitute a counterexample to the revisionist interpretation of the belief-desire model of motivation. But they also afford interesting insight into the genesis of desire on the orthodox model. They suggest that desires as identifiable empirical events may be the effect of social and linguistic conditioning, and are not necessarily to be found in its absence. In Sections 2 and 3 below, I detail the extent to which we buy into this conditioning, in the hope that the resulting unflattering portrait may inspire us to sell out of it.

At this point it is tempting to respond by insisting that all agents have desires, even if they don't know what their desires are, or even what a desire is; and to set about trying to prove these claims by showing that their behavior can be explained by postulating the existence of unconscious or unrecognized desires. I argue in Chapter III that the closer we come to demonstrating the ubiquity of desire as a motive for behavior, the closer we come to a blanket metaphorical description of behavior that leaves far behind any pretense of explaining it. If we cannot distinguish between ontological but contingent desires and other motivationally effective causes of action, we cannot be said to have an independent concept of motivation at all.

1.3. Lewis's Revisionism

To see this, consider next David Lewis' revisionist suggestion for solving the problem of radical interpretation, defined as follows: How do we come to know Karl himself - his beliefs, desires, and meanings, as expressed in his language, and also as we might express them in ours, given the physical facts about Karl?[10] Lewis sets up the problem by assuming all the data yielded by Karl as a physical system *P*, and then devising a way to fill in the information about Karl's beliefs and desires as expressed in our language *Ao*, as expressed in Karl's language *Ak*, and the meanings or truth-conditions of his full sentences *M*, given certain constraints. Lewis describes these constraints as "the fundamental principles of our general theory of persons. They tell us how

[9]B. A. Farrell, "The Criteria for a Psychoanalytic Explanation," in D. Gustafson, Ed. *Philosophical Psychology* (New York: Doubleday, Inc., 1964).

[10]David Lewis, "Radical Interpretation," *Synthese 23* (1974): 331-44. Reprinted in *Philosophical Papers, Volume I* (New York: Oxford University Press, 1983), pp. 108-121.

beliefs and desires and meanings are normally related to one another, to behavioral input, and to sensory input"(111). The problem of radical interpretation, on this account, is the problem of how to read another persons' physical states, events, and behavior, all of which are guided and governed by her understanding of the world, in the terms that define our understanding of the world, without subjective bias or distortion. Do we, and can we, ever really understand what anyone else says or does on her terms? Or must we invariably appropriate her behavior to our own agendas and preconceptions?

That Lewis is to be identified as a revisionist desire theorist (albeit a somewhat tentative one) can be inferred from his stated intention to limit discussion of Karl's propositional attitudes to his system of beliefs and desires "in the hope that all others will prove to be analyzable as patterns of belief and desire, actual or potential; but if not, whatever attitudes resist such analysis also should be included in *Ao* and *Ak*"(109). But Lewis' revisionism becomes much more militant later on, when he asserts that

> [w]e are within our rights to construe 'desire' inclusively, to cover the entire range of states that move us .... Humeanism understood in this inclusive way is surely true – maybe a trivial truth, but a trivial truth is still a truth.[11]

(We are, of course, *within our rights* to construe words in any way we like, as Humpty Dumpty observed.) In this more recent discussion, Lewis goes on to ascribe to a hypothetical Anti-Humean opponent a very odd view: that some desires are beliefs, namely those which are necessarily conjoined with beliefs. More specifically, beliefs about what would be good are claimed to necessarily entail desires for that good. Since Lewis does not cite any philosopher who holds this view or defends it at any length, it is difficult to evaluate its plausibility or its merits.[12] On the face of it, it would seem to have very few. Lewis claims, on behalf of his hypothetical Anti-Humean, that "[i]t is just impossible to have a belief about what would be good and lack the corresponding desire" (324). But surely one may both sincerely believe that, for example, a fair redistribution of resources to the disadvantaged would be good, and also desire not to redistribute one's own unfair accumulation of resources to the disadvantaged; or believe that a trade-in on a new car would be good, yet desire to retain one's 1956 VW Sunroof Sedan forever; or believe that a life of sloth and self-indulgence would be bad, yet desire to live such a

[11]David Lewis, "Desire as Belief," *Mind 97*, 387 (July 1988), 323-332.

[12]Lewis' Footnote 1 cites only criticism – not "criticism and defense of several Anti-Humean views." A subtle treatment of it is to be found in Mark Platts, "Moral Reality and the End of Desire," in *Reference, Truth and Reality*, Ed. Mark Platts (London: Routledge and Kegan Paul, 1980), 69-82. A very plausible, genuinely anti-Humean version of this argument also not cited by Lewis is offered by S. I. Benn and G. F. Gauss, "Practical Rationality and Commitment," *American Philosophical Quarterly 23*, 3 (July 1986), 255-266.

life. It is very hard to see how any such connection between beliefs and desires could be a necessary one – at least not without begging several important questions.[13]

Lewis' general theory of persons in "Radical Interpretation" both implicitly defines the key theoretical terms of "belief," "desire," and "meaning," and also deploys them "to make an empirical claim about human beings – a claim so well confirmed that we take it quite for granted"(111). The *definition* of a person's system of belief, desire, and meaning requires that that system must more or less conform to the principles of the theory. The *empirical claim* this theory makes is that most human beings in fact have systems of belief, desire, and meaning that conform to the principles of the theory. Since the concepts of belief, desire, and meaning are common property, Lewis reasons, the theory that implicitly defines them "had better ... amount to nothing more than a mass of platitudes of common sense, ... on pain of changing the subject" (112). Thus Lewis, like Goldman, wants to claim both ubiquity and ontological primacy for his concept of desire. What makes him a revisionist, however, is his acknowledgment that belief-desire talk is primarily a theory-laden convention – a convention laden with a universalistic theory.

The constraining principles of Lewis' general theory of persons are as follows.

> (1) A *Principle of Charity* constrains the relation between *Ao* and *P* such that Karl is represented as believing and desiring what we would believe and desire, were we in his place, given the existence of a common inductive method *I* and underlying system of basic intrinsic values *V*, respectively.
>
> (2) A *Rationalization Principle* also constrains the relation between *Ao* and *P*: The beliefs and desires ascribed to Karl by *Ao* should allow us to interpret the gross physical behavior given by *P* as maximizing Karl's utility – which Lewis takes to be equivalent to an interpretation of his behavior as rational. "Thus if it is in *P* that Karl's arm goes up at a certain time, *Ao* should ascribe beliefs and desires according to which it is a good thing for his arm to go up then" (113).
>
> (3) A *Principle of Truthfulness* constrains the relation between *Ao* and *M* such that the beliefs and desires ascribed to Karl by *Ao* should preserve the truthfulness of Karl's utterances in his language *Ak*.
>
> (4) A *Principle of Generativity* constrains the assignment by *M* of truth conditions to the sentences of Karl's language *Ak* to that which is finitely specifiable, reasonably uniform and simple, and conforms to a set of standardized semantic and syntactic rules.

[13]Michael Stocker takes up these questions in his "Desiring the Bad: An Essay in Moral Psychology," *The Journal of Philosophy LXXVI*, 12 (December 1979), 738-753.

(5) A *Manifestation Principle* constrains the relation between *P* and *Ak*, and also *Ao*, by stipulating that Karl's beliefs, as expressed in his own language, ordinarily should be manifest in his dispositions to speech behavior; i.e. Karl is assumed to be truthful, honest, and to possess integrity of thought with utterance. Lewis acknowledges the difficulty of stating a companion Manifestation Principle of the desires in *Ak*, but suggests that there probably should be one; we shall return to this difficulty momentarily.

(6) A *Triangle Principle* constrains the three-way relation among *Ao*, *M*, and *Ak*, by requiring that Karl's beliefs and desires come out the same whether expressed in his language or ours.

Lewis' preferred method for solving the problem of radical interpretation, as he has stated it, has three steps: The first step is to use *P* as a source of information on Karl's behavior and life history of evidence to fill in *Ao*, by means of the Rationalization and Charity principles. The second step is to use *Ao* as a source of information about those of Karl's attitudes pertaining to speech behavior to fill in *M*, in conformity with the Truthfulness and Generativity Principles. The third step is to use *Ao* and *M* to fill in *Ak*, by means of the Triangle Principle. The satisfaction of the Manifestation Principle then follows automatically from that of the Truthfulness, Rationalization, and Triangle Principles, and so is redundant.

But Lewis' solution to the problem of radical interpretation suffers on account of his adherence to the revisionist variant of the desire model of motivation. First, recall the status of the theory of persons of which the six constraining principles are constitutive. They are, by hypothesis, systematizations of our commonsense views on the empirical interrelations of belief, desire, and meaning in most people. As we have just seen in considering the Tabwa of West Africa, this hypothesis by itself is controversial. The problem with the Principle of Charity is its presupposition of the belief-desire model in the first place. Despite Lewis' stipulation of *V* in (1), this model leaves no room for basic intrinsic values, because it locates the source of value in individual and subjective desires. Individual desires held by different subjects at different times may fortuitously coincide in conferring value on some common state of affairs, such as food or shelter, with some degree of statistical regularity – on which its value is contingent. But its *intrinsic* value would require that its value be self-conferring (this is what the word "intrinsic" means), not by its statistical occurrence as an object of subjective desire.

But when we examine the statistical occurrence of various states of affairs in the subjective desires of various individuals, we see that not only are the resulting values not intrinsic; they are, for the most part, not even held in common. It is far from obvious that, for example, the Principle of Charity

obtains between persons of neighboring voting districts, much less communities, cultures, societies, or nations. The prevalence of xenophobia on the global as well as local scales calls into question the extent to which others can be supposed to share with us a common set of basic values (let alone intrinsic values), such that they can be represented as believing and desiring what we would believe and desire were we in their place. Instead, the Principle of Charity would seem rather to hold primarily within relatively small and insulated communities, within which conventions of diplomacy and prohibitions against broaching the topics of politics, religion, or sex are superfluous to maintaining at least the appearance of mutual understanding.

Of course the scope of the Principle of Charity could be extended to cover other cases in which such conventions and prohibitions are essential, by sufficiently weakening the qualifications for something's counting as a common system of basic values – of an extrinsic and contingent sort; but then even to state the Principle of Charity as a constraint on our general theory of persons would be otiose. The insufficiency of the Principle of Charity to the facts of social diversity undercuts not only Lewis' assumption of a shared fund of basic beliefs and desires; but thereby the revisionist assumption of the ubiquity of desire more generally.

Similarly with the Principle of Rationalization. I argue in Chapter III that there are, in any case, internal, structural problems with this model that constrain its meaningful empirical application. But even were those problems not to arise, it would seem simply to be false empirically that most people maximize utility in their physical behavior, even if that claim is sufficiently weakened by qualification to bring it dangerously close to vacuity. For even if we suppose a rational agent always to act on those beliefs about how best to satisfy her desires, *given* her incomplete knowledge of probabilities and available resources, inadequate computational skills, distorted judgment, lack of mobility, disinclination to reflect seriously on her true desires, and so on, it is still doubtful whether most agents turn out to be rational, even in this attenuated sense. Neurosis, criminal insanity, weakness of will, and impulsive behavior provide abundant counter-evidence to the empirical applicability of Lewis' Principle of Rationalization. Of course this is not to deny the value of this principle in a possible *normative* theory of persons (at least not in this discussion). It is merely to restrict its role to the definitional. We may simply stipulate its applicability to some, sufficiently idealized community of human agents, but we should not expect its routine empirical confirmation in the behavior of actual ones. The insufficiency of the Principle of Rationalization to the facts of human imperfection undercuts not only Lewis' good-faith assumption of shared instrumental rationality; but thereby the applicability of the Humean model of instrumental rationality more generally. Both of the Humean conception's models – of rationality as well as of motivation – are

inadequate to the psychological facts, and ensure the inadequacy of Lewis' first two principles accordingly.

However, without the empirical applicability of these first two principles, it is hard to see how they might enable us to carry out the first step of Lewis' proposed method for solving the problem of radical interpretation. If they fail adequately to conform to the empirical facts, then it is hard to see why we should use them to constrain the relation between *P* and *Ao*. To fill in *Ao* from *P* as constrained by the Principles of Charity and Rationalization would require us to ignore too much of the physical data actually supplied by *P*. The result would be an interpretation of Karl's beliefs, desires and utterances that would be inadequate to explain them; and we would thereby lose the use of *P* as an independent source of information about Karl. *P* would be relevant only in so far as it could be made to confirm the two Principles – a Procrustean task indeed, if the counterexamples just described are sufficiently pervasive. This would be to subordinate *P* to the constraints of the two Principles plus *Ao*. Then although the second and third steps might yield an interpretation of Karl's beliefs, desires, and sentences in *Ao*, they would not be the *radical* interpretation that Lewis originally set out to give; for it would not enable us to understand Karl's behavior on his terms in our language. Rather, it would require us to construct, from the data supplied by *P*, an interpretation of Karl's behavior on *our* terms, in our language. It would thereby eliminate by fiat all the data that make Karl distinctively different from us, and therefore an object of interest or curiosity, in the first place. In order to understand Karl on *his* terms, we must be willing to modify, abdicate or add to some of our own in light of them. So faithfulness to the physical data supplied by *P* requires us to reject or revise the Principles of Charity and Rationalization, whereas faithfulness to the latter requires us to reject salient portions of the former.

But now suppose we modify the first two principles, and with them their underlying models of motivation and rationality respectively, by weakening them to fit the data. Suppose we accept vacuous attenuations of them that make it true by definition that, first, Karl is represented as believing and desiring what we would believe and desire, were we in his place, given suitably weakened criteria for what it might mean for us to be in his place, to share a common set of basic values, and so forth; and, second, Karl is represented as maximizing utility in his behavior, given suitably weakened criteria for what counts as maximizing utility, of the sort that often discussed under the rubric of revealed preference theory and that will be examined at length in the following chapter. Suppose, that is, that we interpret the physical data yielded by *P* so that they fit the Principles of Charity and Rationalization by stipulation. On this assumption, there is no further question to be asked about the empirical applicability of this theory of persons, for it closes the gap between what we take to be the case and how what we take to be the case is to

be theoretically interpreted. In this case, we gain nothing by invoking the two Principles in order to explicate our general theory of persons, because they are both trivially confirmed by any and all instances of Karl's behavior, whatever it may be. That is, the Principles of Charity and Rationalization add nothing of substance to our interpretation of the data furnished by *P*. They enable us to conceptualize them, but not properly to understand them. Rather than subordinating *P* to the two Principles plus *Ao*, we in this instance subordinate the two Principles plus *Ao* to *P*. Hence while our rendering of Karl's beliefs, desires and utterances in *Ao* and *Ak* may be described as radical in its assimilation of the physical data supplied by *P*, it is not, properly speaking, an *interpretation* of that data.

The problem, it would seem, is Lewis' reliance on the Principles of Charity and Rationalization to motivate his solution. These are the culprits because they alone mediate the relation between *Ao* and *P* on which the formulation of the problem of radical interpretation depends. In interpreting Karl's behavior in *P* in terms of what we would believe and desire were we in his place, we reject the factual basis of cultural xenophobia – namely the inherent subjectivity and contingency of desire, in order to embrace the fiction of cultural appropriation. The comforting supposition that Karl really is basically just like us, except for relatively superficial differences in, say, life history, background, or appearance may be seen as a well-intentioned antidote to cultural xenophobia. But this antidote is purchased at the price of acknowledging information about the very real empirical divergences between Karl's motivational states and ours that make the problem of radical interpretation an important one. This, in turn, ensures the inaccessibility to us *in theory* of Karl's divergent motivational states themselves - and our continuing, guileless perplexity at the chasms of mutual incomprehension that can be generated by the introduction of sensitive topics into otherwise innocuous conversation. Thus our charitable impulses in trying to understand Karl's motivational states backfire, by rendering them yet more elusive of our attempts.

Similarly, in interpreting Karl's behavior given by *P* as invariably utility-maximizing, we compound the elusiveness of Karl's motivational states by the elusiveness and ubiquity of our own. In order consistently to apply the Principles of Charity and Rationalization conjointly, we must suppose the latter to describe our own behavior as well as his. We must interpret our own behavior, too, as invariably and ubiquitously maximizing utility, even in apparently irrational behavior; this is the essence of the revisionist variant on the desire model of motivation. As I argue at greater length in Section 3 below, this interpretation requires us to ascribe to ourselves unknown, hypothetical final desires that our actual behavior, however suspect, efficiently satisfies. Thus the ubiquity of the utility-maximization hypothesis renders our own motivational states just as obscure and elusive as Karl's, for

the Principle of Rationalization requires the stipulation of final desires to which our behavior is instrumental that are *in theory* inaccessible to us.

But these two features – elusiveness and ubiquity – are precisely what obscure the data supplied by *P* to our attempts to understand them. The Principle of Charity ensures the elusiveness of *P* by expropriating it to an account of our own behavior in *Ao*. The Principle of Rationalization then extends the scope of that account, by stipulating ubiquitous but inaccessible desires to which that behavior is instrumentally efficacious. This renders *P* as a source of information about *Ak* and *M* not just practically but theoretically mysterious – like a noumenal thing in itself, whose sole function is to represent our knowledge of Karl as inherently self-limiting.

Now we are in a better position to see why it might be difficult to formulate a companion Manifestation Principle for desire in Lewis' general theory of persons. Such a principle would seem to have to run something like this:

> (5') Karl's desires, as expressed in his own intentional actions, ordinarily should be manifest in his dispositions to gross physical behavior.

However, the attenuated Principles of Charity and Rationalization preclude appeal to Karl's gross physical behavior itself as an *independent* source of information about Karl's actions and desires, for it is already contained in *P*. Like *P*, then, that behavior is admissible only to the extent that it conforms to the two Principles – i.e. only to the extent that it supports the definitional part of the theory. This means that what is to count as intentional action, and what intentional action is understood to have been performed, is also determined solely by the definitional aspect of the two principles – which thus determine what is taken to be manifest in Karl's dispositions to behavior.

Now if we think of language use as a special case of intentional action, and speech behavior as a special case of gross physical behavior, then on the Humean conception, speech behavior satisfies the speaker's desire to express her beliefs. So the redundancy of a Manifestation Principle for desire implies the redundancy of the Manifestation Principle for belief, independently of the Triangle and Truthfulness principles (Lewis' third step, above). To be sure, once we assume the Manifestation Principle for Desire, the companion principle for belief is unproblematic, for we thereby implicitly assume criteria for something's counting as speech behavior, and so, implicitly, for something's being expressable in a language. But without the criteria of speech behavior and language implied by the Manifestation Principle for belief, it is difficult to see what use we might make of the Principle of Truthfulness, nor how the Triangle Principle might fail to be satisfied. And it is unclear how we might derive such criteria without prior independent

criteria of Karl's intentional action in general. But since such criteria are exhaustively specified by *P* subject to the constraints of the Principles of Charity and Rationalization, no further independent criteria are to be found.

It appears, then, that neither the revisionist nor the orthodox analysis of desire is fully satisfactory.[14] What is right about the revisionist view is that it regards belief-desire talk primarily as a theory-laden convention, rather than as referring to ontologically basic events that all human beings must be supposed to experience. What is wrong with it is the effect of what we might fancifully describe as its cultural imperialism: This view is problematic because its universalistic ambitions render the concept of desire both vacuous and bereft of explanatory force. What is right about the orthodox view, on the other hand, is that it regards belief-desire talk as referring to particular

---

[14] Simon Blackburn's resourceful attempt to salvage the belief-desire model in his *Ruling Passions* (New York: Oxford University Press, 1998) came to my attention too late to discuss in the text. In it he proposes a similar variation on Davidson's principle of charity that attempts to turn the vacuity of the belief-desire model of motivation into a virtue. According to API,

> [i]t is analytic that creatures with beliefs, desires, and other states of mind, behave in ways that (best) make sense (and not in ways that make no sense), given those states of mind (55) … We know what a desire is by knowing what it would make sense to do in the light of having the desire; but then we know whether someone has the desire by seeing if this light is one that makes good sense of what they do. API can be true because desires and beliefs are defined by what it is that they make sense of. But they are attributed by what they make people do, under the rubric that people do what makes sense to them (58).

According to API, we understand a desire through the action that would be rational if one had it, and attribute the desire that makes the corresponding action rational. Thus on Blackburn's account, agent behavior satisfies both a norm of rational intelligibility on the one hand, and also a descriptive causal explanation on the other. However, API does not state an analytic principle. API implies that by definition, an agent performs just those actions that best illuminate, either to herself or to others, the beliefs and desires that motivate it. Construed as an account of rational motivation it would seem to be, at the very least, *non*-trivially true in a limited range of cases, in so far as it stipulates, in addition to the operative beliefs and desires, a further desire to make those beliefs and desires intelligible to the observer, whether oneself or another. Construed as an empirical causal description, API would seem equally non-analytic, because susceptible to falsification by any such observer to whom the motivationally effective causes of agent behavior are generally cryptic, mysterious or otherwise irrational – which in most real-world adult negotiations would seem to be the rule rather than the exception. Were API to have some such implications neither for first-person rational motivation nor for third-person causal attribution, it would be empty without being vacuous. Thus within that limited range of cases in which agents do, in fact, behave in such a way as to make intelligible rather than mysterious the desires we ascribe to them, this would seem to presuppose the same sort of shared assumptions and mutual coordination that, I have argued above, Lewis' revisionism also presupposes.

discriminable events the occurrence of which are contingent on other events. What is wrong with it is that it requires us to construe desires as ontologically basic events that all human beings experience, rather than as more localized products of our social and linguistic conventions. What is needed, it seems, is a view that accounts both for the contingent experiential reality of desires as genuine motivational states, and also for their ontological obscurity.

2. Desire and Externality

2.1. A Representational Analysis of Desire

I now offer an alternate account of what a desire is that attempts to do just that. This account formalizes somewhat the conception of desire I deploy in Chapters V through XIV below, makes explicit some of the presuppositions of Richard Brandt's "Kantian" account of desire I discuss in Chapter XI, and elaborates some of the commonsense connotations of the term as we tend to use it in ordinary discourse. It also sheds some light on certain familiar afflictions that beset the belief-desire model that neither the orthodox nor the revisionist variants alone can explain. Most important, it elaborates some of the psychological and characterological implications of the belief-desire model of motivation.

Define a *desire* as a three-place relation between a conscious subject *S*, an object, event or state of affairs *O*, and conscious and occurrent phenomenological representations R_1 – R_3, such that

(a) *S* has R_1 of *O*, such that R_1 represents *O* as lacking (i.e. wanting) in *S*;

(b) *S* has R_2 of *S*, such that R^2 represents *S* as lacking (i.e. wanting) *O*;

(c) R_1 and R_2 conjointly cause *S* to feel discontent, anxiety, insecurity, and craving for *O*;

(d) *S* has R_3 of *S*, such that R_3 represents *S*'s acquisition or achievement of *O* as causing *S* to be whole and sufficient relative to *O*;

(e) R_3 causes *S* to anticipate feeling satisfaction, gratification, security, self-sufficiency, and/or fulfillment.

Call this the *representational analysis* of desire. Representations themselves, unlike assertions, do not necessarily have a propositional structure. Unlike beliefs, they themselves do not necessarily have the complex structure of an intentional attitude. Unlike images, they are not necessarily visual. Like thoughts, representations are necessarily conceptual. But they are both more psychologically atomic and more general in content than any of these.

The representational analysis of desire is sympathetic with the orthodox variant. It similarly conceives a desire as a contingent, occurrent empirical event. It moves beyond the orthodox variant, however, in explicitly insisting

on the phenomenological and representational – and therefore conventional and theory-laden – nature of a desire as an object of conscious belief. This means that the occurrent empirical event that the term "desire" picks out is both the object of and a reaction to theory-laden thoughts and concepts. On this analysis, if a motive is not conscious – i.e. not systematically and causally interconnected with certain representations, it needs to be given another name in order to distinguish it from those which are. This stipulation will help us to see in Section 3 below some of the conceptual pitfalls to which the orthodox and revisionist variants of the belief-desire model succumb.

On this analysis, in order to be said to have a desire, one must believe of oneself that something is wanting, and react accordingly. So a necessary condition of a conscious motive's or drive's being a desire is that the subject must be able to represent to himself the object of desire *O* as, strictly speaking, a *want* in the subject which the object represented will supply. So what one desires depends on one's self-conception, on what one conceives oneself to lack. This stipulation implies the possibility of distinguishing between what one conceives oneself to lack – i.e. what one desires; and what, from a third-personal and distanced perspective, one can be said to lack in fact. So it allows for a distinction between want and objective deprivation, and suggests that there is no necessary connection between the two.

This stipulation also enables us to distinguish desires from other conscious motives or drives such as intentions, resolves, impulses, obligations, compulsions, and whims. It captures the defining content of a desire, namely the representation of a designated object, event or state of affairs as absent or lacking, such that it must be supplied in order to restore one's self-conception as sufficient or complete. This defining content may be found in mundane, ordinary objects of desire such as the consumption of a jelly doughnut or the purchase of new windshield wipers, as well as in more central and pervasive ones such as the desire for a simpler life, or for meaningful work, or for moral goodness.

This representational analysis fills a significant lacuna in Brandt and Kim's definition. On their account, one just does feel satisfaction or pleasure in thinking about the object of desire, and frustration in its nonattainment: these are the occurrent internal events that cause one to pursue it. But why one should have these strong mental and emotional responses to a particular object remains a mystery. According to Brandt and Kim, desires are arbitrary in a sense well-captured by Thomas Nagel's designation of them as "unmotivated desires:" the orthodox variant offers no further rule-governed psychological explanation for their occurrence. Rather, we must recur to physiology or brain chemistry if we wish to trace the causal chain any further back. But this seems incomplete. Desires, as liberal sociologists are quick to assure us, are not as arbitrary as all that. The representational analysis provides the missing link. One feels satisfaction in entertaining a particular

object of desire, and frustration in its nonattainment, because one represents it to oneself as lacking in one. And one is moved to obtain that desired object because one further represents it to oneself as restoring one to wholeness and sufficiency: Once I have *x*, we tell ourselves silently, I'll be fine. (And, for an instant, I am.) These representations in turn can be explained by forces of socialization, acculturation, and familiar forms of consumer indoctrination such as mass media marketing and entertainment.

The representational analysis of desire is not susceptible to Michael Smith's objections to what he calls the "strong phenomenological conception of desires."[15] In particular, it does not state that an agent has a desire only if he believes that he does; nor, therefore, does it imply that an agent knows infallibly what he desires. Rather, it states that the agent has a desire only if he represents something as an object of desire and reacts accordingly. This condition is consistent with his having beliefs about what he desires that turn out to be false. So, for example, Oscar might represent a charbroiled steak as an object of desire - i.e. as lacking etc.; and react accordingly – i.e. with discontent, anxiety, craving, etc. for one. These conditions conjointly imply that Oscar desires a charbroiled steak. Yet they are consistent with Oscar's believing that he desires, not a charbroiled steak, but rather merely a source of complete protein that, among the available options, only the charbroiled steak, unfortunately, can supply (we might suppose Oscar to be an ambivalent vegetarian with a talent for rationalizing his deviations).

Nor is the representational analysis susceptible to Smith's second objection to the strong phenomenological conception, that it "cannot explain how it is that desires have propositional content" (48). Now here Smith makes an assumption about the necessity of propositional content to intentional states that I cannot address adequately until Volume II. Suffice it to say that if it were true that desires *had* to have propositional content, the representational analysis of desire *would* have no trouble explaining this. We would simply reformulate clauses (a), (b), and (d) as intensional belief statements; and clauses (c) and (e) as their corresponding hypothetical indicatives (notice that even this reformulation would not make the representational analysis susceptible to Smith's first objection). These reformulations would expose the propositional content of desires without making the further claim – which I argue in Volume II is much too strong – that the propositional content of those desires which have it exhausts the content of desire overall.[16]

---

[15]Michael Smith, "The Humean Theory of Motivation," *Mind 96* (1987), 36-61.

[16]Smith goes on to argue that desires need not have any phenomenological content, that they need not be felt, that they are best understood as dispositions to behave in certain ways, and that to have a goal just is to desire. In these arguments he does not refer to the early work in action theory that has been done on these issues by Brandt and Kim,

So the representational analysis of desire is inherently self-reflective, but not infallibilistic. It makes of a desire not merely a raw, empirical mental event, but rather a series of conscious mental events that is the product of a certain representational conceptualization by the subject of her inner experience. The representations in question may be conceptual, linguistic, imagistic, kinaesthetic, or some combination thereof. If one does identify and represent one's occurrent mental state as a state of desiring, then one believes, either dispositionally or occurrently, certain propositions about oneself, for example that one is wanting in some respect; that one can be made whole by acquiring or achieving that which one believes is wanting, etc.[17] To accept the belief-desire model of motivation, on this analysis, is then implicitly to believe of oneself and others that we are motivated to action – and so actually *to be* motivated to action – solely by thoughts about what we lack, by beliefs about respects in which we are wanting, to achieve sufficiency and wholeness through the acquisition or achievement of those things. This analysis does not claim that we are motivated in a certain way merely because we think we are. Rather, it claims that our thought that we are wanting in some respect causes certain reactions in us that in turn motivates action to replenish those wants.

Brandt, Dennett, Goldman, Hempel, Kenny, Melden, Pritchard, and Ryle, among others.

The representational analysis of desire also avoids Pettit and Smith's objections to a very strong view that claims desires always to be in the "foreground," i.e. to figure as objects of self-conscious deliberation in an agent's motivational states (Philip Pettit and Michael Smith, "Backgrounding Desire," *The Philosophical Review XCIX*, 4 (October 1990), 565-592). However, I know of no philosopher who holds this view, and Pettit and Smith cite none. They also seem to follow Mill in conflating what one desires with what one believes to be desirable, and overlooking the possibility that one might have a motivational aversion to what one recognizes as desirable. I. L Humberstone ("Wanting as Believing," *The Canadian Journal of Philosophy 17*, 1 (March 1987), 49-62) also makes this mistake.

[17]This interpretation finds its practical analogue in an idiomatic expression current in the world of finance. One says of a brokerage firm that protects an investor against financial losses incurred by delays in executions of stock transactions or other technical problems that it *makes him (or her) whole*. This means that the investor is compensated financially by the firm for any loss resulting from such problems. See Virginia Munger Kahn, "Brokers Making Amends for Trading Problems," *The New York Times* (Sunday, November 2, 1997), Money and Business Section, 8. More generally, to compensate someone for a loss they have incurred is to "make them whole," as in the following example: "Accusing the government of reneging on a promise to make them whole financially for loved ones lost in the September 11 terrorist attacks, several relatives of the dead and injured now say the Justice Department is victimizing them a second time with its tight-fisted handling of the federal compensation fund" (Ralph Ranalli, "Victims' kin decry formula for Sept. 11 compensation fund," *The Boston Globe* (January 14, 2002), A1).

This analysis implies that we perceive the external world as inherently superior to ourselves; as not only malleable through action in the service of our wants, but thereby as a source of gratification of them; as a set of resources for reinstating the wholeness or sufficiency of the self, in which a condition of abundance is transferred from the external world to oneself through one's action. So on the Humean conception of the self, our relation to the external world is one of felt privation. We are motivated to perform some particular action by the promise of restoring the self to wholeness in a certain respect – the respect defined by the desired object we represent to ourselves as lacking. To say, then, that the Humean conception of the self defines and identifies the self by its desires is to conceive of the self as defined and constituted by its self-perceived deficiencies, and its desired objects as external sources of replenishment of these deficiencies.

This analysis also implies that we perceive the external world through the lens of our wants, i.e. as a source of respects in which we are lacking, wanting, or insufficient. All external states of affairs are implicitly evaluated and graded with regard to their suitability as instruments, resources, or approximations of objects of desire, such that the higher the desire-satisfaction rating of a particular state of affairs, the greater its perceptual salience for the subject. This is the essence of egocentrism. Since every state of affairs is assessed according to this criterion, no state of affairs is neutral with respect to it. Different desires may give different colorations and ratings to the same state of affairs at different times, depending on whether it is perceived as an opportunity or a setback relative to one's desires at that time. To the extent that a state of affairs is gradable neither as opportunity nor as setback, neither as attraction nor aversion, it effectively fails to exist for the Humean self. A state of affairs that bears no relation to the defining evaluative function of the self, i.e. desire, bears no relation to an egocentric self at all.

So to perceive the external world through the lens of one's wants is to perceive a world considerably constricted by them. Perceptual salience does not, of course, imply perceptual veracity; precisely the opposite in this case. The primacy of desire-satisfaction as a criterion for evaluating states of affairs as enhancements of, obstacles to, or approximations or embodiments of objects of desire distorts perception of those states of affairs, by magnifying those properties that satisfy or violate the criterion and miniaturizing those which are irrelevant to it. The overriding desire for sufficiency and wholeness leads the Humean self to perceive the external world as a box of tools, instruments, and missing parts; and to ignore or devalue whatever lies outside it.

The representational analysis of desire generates a terminating criterion of rationality for proliferating orders of desires in the Humean conception that, as we see in Chapter VIII.2 below, Frankfurt's concept of second-order desires as regulative of first-order ones is unable to provide. This criterion is

the highest-order desire for sufficiency and wholeness, i.e. R_3 as embedded in (d) and (e), above. This desire *terminates* the infinite regress of orders of desire because any desire, including this one, is by definition an instantiation of the highest-order desire R_3 for sufficiency and wholeness; and because R_3 neither instantiates nor leaves open the possibility of any yet higher-order desire. It is a terminating criterion of *rationality* because, first – to adapt Nagel's criterion of rationality for present purposes, wholeness and sufficiency are ends that can serve as justificatory reasons for actions taken to achieve them; and second, this criterion enables us to evaluate the rationality of any desire – including R_3 itself – by asking whether satisfying it does, in fact, restore the agent's sense of sufficiency and wholeness. – That's the good news for the Humean model of motivation.

The bad news is that no desire can satisfy this criterion, for the reasons Hobbes was the first modern Western philosopher to note:

> [T]here is no such ... *summum bonum*, greatest good, as is spoken of in the books of the old moral philosophers. ... Felicity is a continual progress of the desire, from one object to another; the attaining of the former, being still but the way to the latter. The cause whereof is, that the object of man's desire, is not to enjoy once only, and for one instant of time; but to assure for ever, the way of his future desire. ... So that in the first place, I put for a general inclination of all mankind, a perpetual and restless desire of power after power, that ceaseth only in death. And the cause of this, is not always that a man hopes for a more intensive delight, than he has already attained to; or that he cannot be content with a moderate power: but because he cannot assure the power and means to live well, which he hath present, without the acquisition of more.[18]

The reason no desire-satisfaction can meet the terminating criterion of rationality – that it restore one's sense of sufficiency and wholeness – is that any desire-satisfaction automatically generates a further desire – i.e. a represented want (or lack or insufficiency) – for the satisfaction to continue; and the satisfaction of this further want in turn generates yet a further want to acquire sufficient power to protect the power one already has to satisfy that one. So although it is true that once I acknowledge my desire for wholeness and sufficiency there is no higher-order desire I can have in terms of which that one can be evaluated, it is also true that that highest-order desire for wholeness and sufficiency itself generates an infinitely proliferating series of lower-order wants: for continuance, and for protection and proliferation of the means for continuance, that "ceaseth only in death;" and so prevents wholeness and sufficiency from being achieved.

---

[18]Thomas Hobbes, *Leviathan*, Ed. Michael Oakeshott (New York: Collier Books, 1977), Chapter 11, "Of the Difference of Manners," 80.

One paradigmatic real-life example of Hobbes' observation would be the American growth economy. Unlike a maintenance economy, the American growth economy is driven, in part, by the desire of investors for a high return on their investments (not merely, as is often claimed, by consumer demand, population growth, or the needs of the labor force). Since different investors enter the market at different times, whatever the share price at which investors enter a particular market, they exert pressure on the relevant businesses to increase that price. Thus no share price can be high enough to satisfy investor desire for a high return. One conventional way in which a business responds to this pressure is by increasing its revenues from the product or service it sells. So just as no share price can be high enough to satisfy investor desire for a high return once and for all, similarly, therefore, no revenues can be large enough for the business thus pressured. One conventional way in which a business increases its revenues is by creating new rationales for raising its prices and new incentives for consumers to purchase its products or services. So just as no share price can be high enough and no revenues large enough, similarly no product or service price can be high enough and no quantity of sales large enough to slake investor desire. Since this system of perpetual serial pressures to increase dollar amounts is independent of the actual needs, desires and natural limitations of consumers (including, of course, investors themselves), they are continually inundated as a matter of course by "new, improved" product models, increasingly overloaded with irrelevant or confusing bells and whistles, that are regularly and rapidly introduced, withdrawn, and soon re-introduced with yet more "improvements." Consumers are also pressured into further consumption by businesses that refuse to support the products they sell for longer than the short time span they are on the market, or interlocking businesses that form a monopoly to coerce consumers into "upgrading" each product in order to use others on which it depends. Thus satisfaction of investor desire pressures businesses to maximize revenues and consumers to maximize consumption irrespective of the actual needs or desires of either. Simultaneously, serviceable and reliable products are discontinued because they pull in a steady and predictable profit rather than an escalating one that satisfies investor desire for a larger return on investments. Thus other things equal, share prices, revenues, and sales escalate at roughly the same accelerating rate at which new goods and services are introduced, forced down the consumer's throat, withdrawn, and re-introduced – accompanied by increasingly assaultive and invasive marketing techniques that rely on sex, violence, and disparities in social status to sell irrelevant goods that consumers neither want nor need and for which businesses fail to create a sufficient demand (since no quantity of demand would be sufficient). Hobbes' "perpetual and restless desire of power after power" leads inevitably to fulsome overload.

This case makes it easy to see how desire on this analysis is distinct from pleasure. I argue at greater length in Chapter VI that desire-satisfaction is not necessarily pleasurable; and indeed – *contra* Brandt and Kim, that desire-satisfaction and pleasure are entirely independent of each other. A feeling of *pleasure* is a reaction to a stimulus that causes in one feelings of sensory and/or emotional well-being, happiness, ecstasy, or joy. In one for whom a sense of wholeness and sufficiency is lacking, a pleasureable stimulus can create in one a desire for those feelings, under the misapprehension that these feelings will restore to one a sense of wholeness and sufficiency. But clearly, these two sets of responses are similarly independent of one another, and often not even contingently conjoined. It is possible to experience simultaneously feelings of wellbeing, happiness, ecstasy, and joy on the one hand; and feelings of insufficiency and deprivation on the other. Indeed, some accounts of religious ecstasy postulate a necessary connection between it and a belief in one's own insufficiency. In this case the feeling of pleasure may be inflected by feelings of being undeserving or presumptuous or base; or by anxieties about future pain, punishment or retribution; or by unexpected revelations of one's value or entitlement to pleasure, etc. Or it may be simply and straightforwardly unsatisfying, as too much sugar, sex, socializing, or status often is. There is no real mystery as to how such things (to name only a few) can give both pleasure and dissatisfaction simultaneously. The more important and less obvious point is that any object of desire may fail to give pleasure, and any pleasure may fail even momentarily to satisfy desire.

2.2. A Representational Analysis of Aversion

Next consider aversion. Somewhat analogously to desire, an *aversion* can be representationally analyzed as follows:

(a') *S* has R_1' of *O* such that R_1' represents *O* as overloading *S*;

(b') S has R_2' of *O* such that R_2' represents *S* as overloaded by *O*;

(c') R_1' and R_2' conjointly cause *S* to feel overstimulation, repulsion, disgust, apprehension, fear, and/or pain with regard to *O*;

(d') *S* has R_3' of *S*, such that R_3' represents *S*'s obtainment or achievement of *O* as causing *S*'s wholeness and sufficiency to be attacked, threatened, invaded, overwhelmed and/or undermined by *O*;

(e') R_3' causes *S* to anticipate feeling discontent, anxiety, insecurity, and craving for the eradication of *O*.

(f) *S* has R_4 of *S*, such that R_4 represents *S*'s eradication of *O* as causing *S* to be restored to wholeness and sufficiency;

(g) R_4 causes *S* to anticipate feeling satisfaction, gratification, security, self-sufficiency, and/or fulfillment.

This representational analysis of aversion is not exactly the converse of the representational analysis of desire, because aversion, on this account, is not simply a negative desire, i.e. a desire not to have something that someone else might have a desire for. An aversion is a complex emotion that also includes substantive spontaneous feelings – of distaste, of feeling revolted or oversensitive or invaded relative to the object, for example – that respond to the thought or representation of the object. They then in turn generate further feelings in response to the thought of actually realizing the object, and a subsequent desire to rid oneself of it. So the representational analysis of aversion embeds such a desire in (e'-g). But this subsequent desire is only one part of an aversion. It is not equivalent to it. Aversion also includes quite distinctive and visceral feelings of sensory overload, oversatiation, discomfort, and anxiety. On this revision of the Humean motivational model, the desire embedded in aversion is still the only source of effective motivation to action. But its prior or concomitant feelings may nevertheless cause spontaneous expressions of aversion, such as feeling nauseous, breaking into a cold sweat, heart palpitations, aggression, flight, or other instinctive behavior. It is the threat to one's sense of wholeness and sufficiency that motivates the desire for the eradication of *O*, not simply the experience of overload *an sich*.

Like desire, aversion on the representational analysis is inherently self-reflective. It involves representations of and beliefs about oneself as sufficient relative to *O*, and so beliefs about what will threaten or overwhelm and therefore destroy one's sufficiency. An aversive object is one that one represents to oneself as invasive, i.e. as violating the boundaries of one's self, as intruding into one's self where no interior psychological space has been made for it. An invasive object is one that by definition causes pain or discomfort. To perceive the world through the lens of those desires embedded in aversions is to perceive oneself as wanting in control over one's environment; as unable to vanquish attacks on one's sense of wholeness; as invaded, fragmented, and overwhelmed by alien external objects which the interior psychological space of the self is too limited, fragile, or crowded to accommodate. So, like desire, an aversion includes negative beliefs about oneself: as powerless, fragile and permeable; as well as negative beliefs about the world as overwhelming, threatening, and invasive. This part of the Humean conception of the self thus supplements and underwrites that according to which the self is defined and constituted by its self-perceived deficiencies. Desires and aversions conjointly create a mutually interlocking and mutually supportive set of assumptions about oneself as constitutionally deficient in various respects; and about the external world as correspondingly resource-abundant in some respects and threatening and overwhelming in others.

With the aid of this account of aversion, we can now discern a second reason why R_3 as a terminating criterion of rationality for the Humean conception is nevertheless incapable of satisfaction. It is a well-known phenomenon that objects of desire begin to lose their luster after they are obtained. This is often attributed to fickleness, shallowness, or inconstancy of character. But actually it is instead implicit in the very structure of desire and aversion. After a desire has been satisfied, the desire or want itself disappears – and with it that which conferred psychological value on its object, leaving nothing for the object that satisfied it to satisfy, and so nothing relative to which it is valuable. A desired job, partner, lifestyle or dessert of necessity seems much less desirable after it is obtained because it stops being the object of one's want and starts being the object of one's surfeit – which is to say one's aversion. At the same time that we want the satisfaction of a desire to continue, we do not want the object of that satisfaction around after it has outlived its usefulness as a satisfaction. So to obtain the object of one's desire is thereby not only to devalue it in the act of obtaining it, but *a fortiori* to transform it into an object of aversion. *Absent any other source of value beside desire*, no object of desire can remain desirable for long after it has been obtained, because no desire can endure after it has been satisfied. The infinite proliferation of lower-order desires in the Humean self is matched only by the finitude of their duration. In this the Humean self is both a bubbling cauldron and a bottomless pit, in which countless desires endlessly form, expand, explode, and disappear.

2.3. Funnel Vision

The belief-desire model of motivation defines the Humean self as *future-oriented*, in that the self finds expression and continuity in setting for itself, in the present, some future, extrinsic desired state of affairs that it can anticipate working to actualize over time.[19] This feature of the Humean self can be regarded as the consequence of tying a dispositional analysis of traits of character to the foundational notion of a desire.[20] To call a person generous or corrupt, on this analysis, is to describe a way she is disposed to act under certain circumstances. But since on the Humean conception of the self, all action is motivated by desires the agent wishes to satisfy, the concepts we invoke to describe a person's character or personality denote certain kinds of desires that person is disposed to try to satisfy under the relevant

[19]This is essentially Bernard Williams' notion of character. See his "Persons, Character and Morality," in A. O. Rorty, Ed., *The Identities of Persons* (Berkeley, Cal.: University of California Press, 1976). It is also consistent with Hume's own analysis of the self in Book I of the *Treatise*, given certain qualifications.

[20]See, for example, Richard Brandt, "Traits of Character: A Conceptual Analysis," *American Philosophical Quarterly 7*, 1 (January 1970).

circumstances. The self then achieves full realization to the extent that it succeeds in satisfying those desires.

Indeed, on the desire model of motivation, objects of desire are by definition external to the self that adopts them, even if they consist in an internal modification of some aspect of self or character. The Humean self is *heteronymous*, to use Kant's term, in that the conditions of its expression are objects or states of affairs that are psychologically and/or spatiotemporally external to the self in its present state. This external relation of the self to its desired objects motivates actions performed in order to appropriate those objects.[21] So the full realization of the Humean self consists in bringing into existence those extrinsic desired states of affairs, and regarding them as newly incorporated satisfaction-states of the self. To become a better person of a specified sort, or to acquire a condominium, or a few moments of peace conceived as objects of desire makes of them psychologically (and perhaps spatiotemporally) distant entities in relation to the present state of the self; entities which the self approaches with a realistic plan, not only for the satisfaction of its desires, but thereby for the appropriation of the objects of those desires. As satisfactions of the self, former objects of desire are then available for recycling as instrumental resources in the service of further objects of desire: The few crumbs I crave today fuel my pursuit of a piece of the pie tomorrow. These further objects nevertheless remain remote from the self in its present incarnation, as they must, in order to provide its structure and present motivation.

The objects of desire to which the self is committed thus provide it with an evaluative perspective on its internal states that is remote without being detached. It is *remote* in that it regards the present internal state of the self from the perspective of a future desired object or state of affairs that the self at present lacks. A remote evaluative perspective on the present state of the self follows from one's identification with those objects of desire that the Humean self is conceived presently to lack. From the perspective of the present state of the self, the future goals to which one aspires may indeed promise both satisfaction and value. But from the perspective of those future goals, the present state of the self must seem unsatisfactory at the very least. It is a feature of any self conceived heteronymously that there is a seesaw between present state- and future satisfaction-perspectives that must tilt both ways without being fully anchored in either. Some degree of remoteness from the present state of the self might, perhaps, be avoided by a being that was fully identified only with its desiring-states and never with their intentional objects.

---

[21] "This relation, whether based on inclination or on rational ideas, can give rise only to hypothetical imperatives: 'I ought to do something *because* I will something else.'" Immanuel Kant, *Groundwork of the Metaphysic of Morals*, trans. H. J. Paton (New York, NY: Harper Torchbooks, 1964), Ac. 441; italics in original.

Such a being might act systematically to satisfy its desires, without, however, representing the objects of those desires to itself. Following Frankfurt's terminology, we might describe such a being as a blind wanton. But it is hard to see such a condition as either possible or, so to speak, desirable for human agents.

The heteronymous perspective on the present state of the self is remote *without being detached* because this perspective remains a personal and subjective one, constricted and defined by the desires, expectations and hopes that simultaneously define that individual self. It regards what I am, have and lack from the point of view of what I want, not from any independent point of view from which what I want itself might be critically assessed. This is something like tunnel vision: We regard our present state of insufficiency from a point further ahead in a temporally linear future, at which sufficiency has been restored by acquisition of the desired object. But since on the Humean conception of the self, we evaluate *all* perceptually available objects, events, and states of affairs from the perspective of their suitability to restore what we lack and not merely those on which we finally settle as objects of desire, the perspective is actually shaped more like a funnel: circumscribed, to be sure, by those states of affairs that satisfy the criterion of perceptual salience – i.e. seen as opportunities or setbacks to varying degrees; but narrowing at the location point of the agent and fanning out to encompass and evaluate the entire array of such possible objects, events and states of affairs within the agent's purview as desirable or aversive, as enhancing or as undermining the agent's wholeness and sufficiency.

So the Humean self is also *egocentric* in the sense that it carves up the internal and external world of actual and possible states of affairs in terms of their satisfaction-potential in the eyes of one particular agent, namely itself. I am motivated to satisfy some desire only if the satisfaction in question is mine. If the desire belongs to someone else, then I am motivated to satisfy it only if I have a further desire I might thereby satisfy: i.e. to satisfy her desire. I argue in Chapter VI.1 that this is not to claim that all the desires I am moved to satisfy are inherently egoistic.[22] I may be moved to satisfy my desire to advance the common good, even at considerable personal disadvantage, by the prospect of advancing the common good, not by that of personal satisfaction. Nevertheless, advancing the common good must be satisfying to me; otherwise *I* have no motivation for advancing it. Thus on this conception of the self, that I merely perceive some state of affairs to best contribute to the common good, or to satisfy someone else's desire, is not sufficient to motivate me to try to achieve it. In addition, I must have a desire to so contribute. For in the absence of such a desire, I have no motivation to contribute.

[22]Bernard Williams also argues this in "Egoism and Altruism," *Problems of the Self* (New York, NY: Cambridge University Press, 1975).

These observations underscore the intimacy of the relation between the self and agency, and so the necessity of identifying the self not just with a certain rational structure, but with a motivational capacity. If I were nothing more than a passively rational contemplator, I could have no self whatsoever. For if I necessarily failed to distinguish, among the ongoing panorama of events, some which I caused to occur, I would equally lack the means of identifying those among my experiences that were caused by something else; I could identify no subject to whom these events were happening. But if I were unable to distinguish myself from the events that happened to me, it is difficult to imagine how I might then distinguish my *self* at all. However, that the self must find definition and expression through action does not imply that the self must be future-oriented, heteronymous, and egocentric. Hence it does not follow from the intrinsic connection between selfhood and agency that the Humean conception of the self is necessarily the correct one.

The hypothesis of a Humean conception of the self and its attendant funnel vision conjointly offer an explanation of why moral conduct as an object of desire has a peculiarly self-directed and narcissistic quality. This may manifest itself in the varieties of self-absorption or contextual insensitivity anatomized in Chapter VI; or in an unusual assertion of will and insistence on the conduct even when evidence of its artificiality, insensitivity or inappropriateness abounds; or in close and regular correspondence between the achievement of the goal of the conduct and feelings of satisfaction in the agent. In all such cases, and many others, we have good reason to speculate that the conduct is driven by desire-satisfaction rather than other moral motives such as duty, compassion, or indignation.

I argue in Chapter VI that all desire-satisfaction is self-interested; and also in Chapter VIII.3.2.4 that narcissism directs the interests of the self toward its own self-image and image in the eyes of others. But we can already see that the self-image in which the Humean self necessarily takes an interest is the self envisioned as whole and sufficient, made so by the satisfaction of desire; that the performance of moral conduct as an object of desire is one among the array of such objects that instantiate that higher-order one; and that all opportunities for satisfying this desire are among those surveyed, ranked and graded by the Humean self. Finally, we can anticipate one of the conclusions of Chapter VIII, that narcissism is not, after all, merely a pathological condition of the psyche, as I have argued elsewhere.[23] It is built into the desire-based motivation of the Humean self.

The hypothesis of a Humean conception of the self conjoined with its attendant funnel vision also offer a partial explanation for the moral phenomenon of *ignorance of oneself as a particular* (I offer a fuller explanation in

[23]"Moral Theory and Moral Alienation," *The Journal of Philosophy LXXXIV*, 2 (February 1987), 102-118.

Volume II, Chapter VIII.5). This phenomenon is a familiar and comical one: Mildred, a Machiavellian social climber, complains bitterly about the Machiavellian social climbers she must contend with daily, and plots to destroy them; Mortimer, the consummate hypocrite and liar, fulminates earnestly to his friends against the evils of hypocrisy and lying, fabricating examples of honesty to prove his points; Maxine and Chester, fair weather friends to all who know them, castigate Archibald's inconstancy and betrayal of them both; Lucille glibly condemns Vernon for his glibness. In all such cases, the agent sincerely holds a moral principle and fails to recognize her own violations of it – indeed, sometimes violating the principle in the act of denouncing violations of it by others. An observer of the scenario wonders how anyone can be so blind to their own faults even while discussing them in the abstract. More generally: How can someone advocate a moral principle on the one hand, and simultaneously exemplify its violation on the other, without being aware of the inconsistency?

Kantian-style explanations to the effect that the agent indulges herself by recognizing the inconsistency and making a just-this-once exception[24] do not go deep enough into the Humean conception of the self. For they assume in the Humean agent a perspective from which the inconsistency is recognizable, i.e. an intellectual and cognitive perspective, detached from the demands of desire-satisfaction and the pull of emotion, that conceives all of the agent's behavior as instances either subsumable under abstract principles or their negations, or with which they are consistent or inconsistent. But a self that is motivated and structured according to the Humean conception allows no room for such a perspective, in which one's actions and character exist in a relation of consistency or instantiation to something more abstract than their effects. We have just seen that such a self is beset by funnel vision. Humeans such as Bernard Williams celebrate this constricted perspective as an index of personal integrity. But we see in Chapter VIII.3.2 that personal integrity does not and cannot require the imprisonment within the personal and subjective perspective that funnel vision expresses. This precludes the stance of detached self-reflection on one's actions and emotions that is so central to a Kantian conception of the self.[25] The Humean self in its pure form is limited to assessing its present state in light of its agenda for envisioned desire-satisfaction, from the envisioned spatiotemporal location of those envisioned satisfactions.

---

[24]Kant, *op. cit.* Note 21, Ak. 424-425.

[25] I develop this point at greater length in "Kants intelligibler Standpunkt zum Handeln," in *Systematische Ethik mit Kant*, Eds. Hans-Ulrich Baumgarten and Carsten Held (München/Freiburg: 2001), in English translation at adrianpiper.com; and in *Kant's Metaethics: First Critique Foundations* (manuscript in progress).

The explanation of ignorance of oneself as a particular in a Humean self is simpler: It is that there is no higher-order principle beyond desire-satisfaction itself, embedded in such a self, that it might recognize itself as instantiating. Identified with its desires, the only available vantage point from which its own actions can be assessed is the remote perspective offered by its future-orientation toward its envisioned satisfactions. This is what it means to say that an agent who satisfies a desire at the expense of prudence or duty has "lost perspective": From the envisioned future point in time at which particular deficiencies are supplied and wants replenished, present and salient abstract moral principles are simply items among the array of externally available resources for achieving this. They are instruments like any others, to be invoked, used, applied, or discarded as needed; and their salience and importance varies accordingly.

Thus the Humean self is rigid in some respects and malleable in others. It is *rigid* in its confinement to the subjective and personal perspective of its agenda for desire-satisfaction, which is a characterological constant. But it is *malleable* in its readiness to adapt opportunistically any principle, any perspective, any situation or resource or state of itself to the achievement of that agenda. For a Humean self, there is no inconsistency in violating moral principle while advocating it, because there is no inconsistency in a strategy that utilizes both advocacy and violation of that which one advocates simultaneously in the service of desire-satisfaction. Such an agent advocates the principle when it is convenient, and violates it when it is convenient; there is no contradiction in the possibility that both may be convenient at the same time. To bring the phenomenon to the attention of the agent herself is to invite detailed explanations as to why her actions do not, in fact, constitute violations of the principle at all, but rather something completely different, required by circumstance. Since the deployment of instrumental means in the service of desire-satisfaction is for her uppermost, her actions really are something different: not violations of her principles, but necessary strategies for restoring herself to wholeness and sufficiency. Who could possibly quarrel with that?

2.4. Attachment and Self-Hatred

The representational analysis of desire implies that from the point of view of what one now wants, what one is and has may look more or less promising, but it can never look evaluatively neutral. And in Section 3, following, I argue that it is a consequence of accepting the belief-desire model of motivation that from the perspective of what one really, deeply wants, what one is and has cannot even look promising, by hypothesis. In either case, one's experience of desiring confers evaluative coloration, not only on the world, but on oneself as one is. One must always size up the world with an eye to its resources for satisfying one's desires, and one's present condition as

one of potential for satisfying those desires. We might say that the funnel vision of the Humean self entails a retrospective personal self-awareness, since from the perspective of one's future satisfactions, one's present condition must always appear inadequate.

Thus there is an elemental sense of inferiority attendant on representing one's self in the terms the Humean conception offers. This is engendered not only by the comparison of one's own state of privation with the plenitude of the external world as a resource for fulfilling it; but even more centrally by the self-evaluation demanded by the funnel vision perspective on one's present wants. By comparison to that envisioned (but unrealizable) future self, made whole by the satisfaction of desire, one's present condition is, in fact and inevitably, inadequate; and so long as one continues to have desires, one can never catch up to it. A felt sense of inferiority is a permanent psychological feature of the self on the Humean conception.

I have argued above that the Humean conception of the self implies that one's reflective view of oneself from the perspective of the objects of one's desires is remote without being detached. But not only is this perspective not detached; it is exactly the opposite of being detached. It is one of deep and obsessive attachment. I shall say that an agent *A* is *attached to* some object, event, or state of affairs *x* if

> (a) *x*'*s* existence is a source of personal pleasure, satisfaction, or security to *A*;
>
> (b) *x*'*s* nonexistence elicits feelings of dejection, deprivation, or anxiety from *A*; and
>
> (c) these feelings are to be explained by *A*'*s* identification with *x*.

And I shall say that *A identifies with x* if A is disposed to identify x *as* personally meaningful or valuable to A.

One can be attached to some *x* without desiring it, for example if *x* is a longstanding authority figure whom one regards with a mixture of respect and revulsion. But to say that one is attached specifically to the object of one's desire *O* is to say that *x* = *O* in (1.a) and (b), above; and that failure to obtain this object of desire causes one to feel dejected, deprived, and/or anxious; discontented, inferior, insecure. And to say that one's self-reflective perspective is one of attachment is to say that one's attachment to the objects of one's desires gives one a highest-order attachment to one's representation of oneself as made whole and sufficient by them (i.e. R_3 in (1.d), above), such that failure of R_3 to reflect an actual condition of wholeness and sufficiency in oneself elicits feelings of dejection, deprivation and/or anxiety. R_3 then becomes an object of desire *O* whose attainment would replace these feelings with different ones: of satisfaction, gratification, security, self-sufficiency, self-confidence, and/or fulfillment. Thus a Humean self has a central attachment

to its terminating criterion of rationality, namely the highest-order desire for wholeness and sufficiency.

In the event that this object of highest-order desire were unattainable, feelings of dejection, deprivation, and/or anxiety would be a permanent character trait of a Humean self. But we have just seen in Subsection 2.2, above, that this object of desire *is* unattainable, because it automatically generates supplementary, lower-order desires for the continuance and protection of the satisfaction of this one that in turn proliferate without limit – thus insuring that feelings of dejection, deprivation, anxiety; discontent, inferiority, insecurity, and craving are, indeed, permanent character traits of the Humean self. Now let us consider some of the more practical implications of this psychological conception.

An agent who is attached, in the sense just defined, to the objects of his desires as replenishments of his variously perceived insufficiencies will experience any failure to satisfy them as increasing his insufficiency. For in failing to satisfy the lower-order desire, he is simultaneously exacerbating his failure to satisfy the highest-order one, and thereby generating further lower-order ones. Each failure of desire-satisfaction thus ramifies throughout the structure of the Humean self, and further expands the range and depth of its sense of privation. And so with every such failure, the felt insufficiency, sense of inferiority, and so the negative self-evaluation of the self from the perspective of its envisioned future increases. But the persisting unattainability of R_3 as an object of desire is identical to the persisting inability to close the gap between one's self and one's fundamental self-conception, i.e. to live up to the ideal that guides and motivates one's actions. Relative to that ideal, the Humean self experiences itself not only as insufficient, but therefore as *de*ficient. Hence every thwarted attempt to satisfy its lower-order desires intensifies its self-dislike to the point of self-hatred. In this way the motivation to succeed in satisfying some desire or other, any desire, so as to restore self-esteem intensifies and escalates as a bulwark against a downwardly spiraling self-hatred. The quest for desire-gratification intensifies as an antidote not only to privation and insufficiency but to self-hatred for the Humean self, and its need for external infusions of esteem increases correspondingly.

By *self-hatred*, I will mean the belief – with its concomitant feelings of revulsion, shame and despair – that one is inferior, to varying degrees, to everything and everyone external to oneself. Thus self-hatred is the subjective expressive counterpart of the Humean self to its representation of the external world as consisting in an abundance of resources for replenishing its felt insufficiencies and restoring itself to wholeness. A self-hating agent perceives himself as inferior to external others who, because they are other than himself, by definition have what he lacks. He perceives himself as inferior to external nonhuman things that, because he sees them as potential resources for desire-

satisfaction, are by definition what he lacks. And he perceives himself as inferior to his envisioned future self whose lacks have been replenished.

Thus self-hatred is an inherently relational, comparative, and quantitative emotion. It is *relational* and *comparative* in that it depends on pairwise comparisons of relative status, such that his own perceived inferior status is perceived as a function of others' perceived superior status, and others' envisioned inferior status – i.e. as lacking what he has – as effecting his own envisioned superior status. It is *quantitative* in that it calibrates the degree of his inferiority to different external others with respect his own gain or loss relative to theirs. From the perspective of a self-hating agent, any gain to another is an aversive loss to oneself; and any loss to oneself is an aversive gain to another. Conversely, any loss to another is a desired gain to oneself; and any gain to oneself is a desired loss to another. So self-hatred is also an inherently *competitive* emotion. An index of the pervasiveness and depth of self-hatred is the degree to which the agent implicitly keeps count: no gain to himself, no matter how much of a desired loss to another it is perceived to exact, suffices to restore equity.

Self-hatred is also an implicitly *envious* emotion, because if a loss to another is equivalent to a desired gain to oneself, then a Humean self is naturally and necessarily willing to suffer loss to itself if this effects loss to another, provided that the gain to itself it obtains through the other's loss outweighs the loss it suffers in order to effect that loss to the other. Envy pays in those cases in which the consequence of one's self-imposed loss is a greater self-directed gain. In Chapter X.3.1.2 below we see that Rawls adopts both a Humean conception of the self and also a stipulation that the parties in the original position are not moved by envy. But we can already see here that these two assumptions are mutually inconsistent. Envy is implicit in the belief-desire model of motivation because one's attempts to achieve the impossible goal of wholeness and self-sufficiency necessitate personal sacrifice when this effects another's loss that maximizes one's own competitive and status-comparative gain.

Thus self-hatred within the Humean conception presupposes belief in a zero-sum hierarchical game in which the goal is to enhance one's own status-superiority and the means is to reduce the status-superiority of others. Because, like all available means, the desire for this goal engenders not only further lower-order desires for them, but in addition desires for the means to protect them, the acquisition and replenishment of means can provide satisfaction independently of that envisioned for the goal. Since this goal is an impossible one for a Humean self to achieve, satisfying desires for means and resources of various kinds provide a more immediate source of gratification that may effectively outweigh and replace that envisioned in obtaining the impossible goal of status-superiority. Thus not only does a Humean self grade and sort the external world into opportunities and setbacks relative to its

wants. More specifically, it evaluates them relative to its instrumental wants, i.e. as means to further desired ends. In practice, then, the highest-order desire for wholeness and sufficiency translates into a pure time-preferential desire for power; and the concept of intrinsic value becomes nugatory.

Self-hatred differs from shame and guilt. It differs from shame in that it presupposes no shared social ideal by comparison with which one regards oneself as defective and so vulnerable to others' ridicule. On the contrary: one perceives oneself as inferior relative to every ideal and every reality, whether social or personal (this is the person who, for example, feels offended at your casual inquiry as to how he is, patronized by your interest and scorned by your indifference). And it differs from guilt in that it involves (like shame) a negative evaluation of the whole person, not merely a single action for which one can be held responsible and required by others to make amends. Self-hatred nevertheless shares with both a negative social dimension, in this case the need to hide one's inferiority from both oneself and from others in acts of self-deception, dishonesty, and hypocrisy.

Self-hatred on the Humean conception also differs from self-criticism, respect for authority, and respect for higher status of any kind. Nor does acknowledgement of one's own lower status imply self-hatred. On the contrary: the ability to acknowledge oneself as flawed, imperfect, or inferior in some respect relative to some valued standard or person who is perceived to meet it presupposes self-esteem, because it presupposes that any such judgment cannot devalue the whole person. Self-esteem thus shares with guilt the ability to isolate and criticize particular actions or characteristics without reducing one's status as a person relative to others in one's own eyes. Because a Humean self permanently lacks a sense of wholeness, any acknowledgement of imperfection reinforces its felt inadequacy.

Because this condition is a permanent and fundamental condition of its agency, the Humean self may experience its self-hatred as natural, neutral, and familiar rather than as traumatic. Indeed, the inherent aversiveness of this condition may be indistinguishable from – in fact, may *be* the motivational spark behind all action in the service of desire-satisfaction. Because all such action is forward-looking with reference to its envisioned future end-state of wholeness and sufficiency restored, action serves to distance the Humean self temporarily from its underlying aversive condition – to which inaction returns it and from which further action is its only escape. So such a self is most clearly marked by the compulsive, continuous, and manic quality of its activity. It conceives and identifies itself through doing rather than being, for whereas its being is poisoned by self-hatred, its doings are fueled by hope. And as its self-hatred increases, so does the attachment to the objects of its

desires as sources not only of satisfaction and wholeness, but thereby of self-esteem.[26]

Thus we may recast the highest-order terminating criterion of rationality for the Humean conception, alternately, as the aversion to self-hatred, i.e. as R_3' in (1.a'-g) above. This aversion would serve as the highest-order causally effective motive for the Humean self, and also as an alternative way of explaining attachment to the objects of one's lower-order desires. Think of the highest-order desire for wholeness and sufficiency and the highest-order aversion to self-hatred, then, not as equivalent but rather as mutually interdependent.

These two mutually interdependent highest-order criteria may explain how it is that thwarted desires may elicit in a Humean self not merely frustration or discontent, but also further desires: for revenge, reparation, or recompense. If a felt failure of wholeness and sufficiency is interdependent with intense feelings of self-hatred to which one is averse, then those feelings will overwhelm and threaten whatever remaining sense of sufficiency the self may retain. Then if some form of desire-satisfaction is not immediately forthcoming, some other form of compensation – some substitute that promises the restoration of completeness – must be. Of course the desire for a substitute for desire-satisfaction is subject to the same frustrations as that for which it is supposed to substitute, since no desire-satisfaction endures, nor fails to generate further dissatisfactions. Persistent frustration of desire, conjoined with the persistent and standing desire for recompense, lead one beyond the object of desire to a persistent and reified sense of oneself as a victim of deprivation and injustice. This not only exacerbates the proliferation of lower-order desires, but rationalizes their pursuit to the agent himself.

It also thereby rationalizes any further infliction on others of deprivation or injustice in turn. We have already seen above that a Humean self regards itself and others as players in a zero-sum game in which the stakes are the accoutrements of desire-satisfaction, namely power and status-superiority, such that losses of these things to others are perceived as gains of them to oneself. For the Humean self, sadistic desires and their resulting acts of spite, revenge, or aggression against others are a natural expression of the self-hatred engendered by attachment to the objects of one's desires. I argue in Chapter VI.3 that a sadistic person takes satisfaction both in others' suffering and also in being the instrument of it; and also that satisfying sadistic desires accelerates self-brutalization.

Finally, the sense of oneself as a victim of injustice and deprivation consequent on the pursuit of recompense for thwarted desire-satisfaction rationalizes unlimited consumption of objects and experiences perceived as satisfaction-substitutes: of commodities for friendship, sex for love, food for

[26]I am grateful to Hans and Linda Haacke for discussion of the concept of self-hatred.

sex, status-superiority for life purpose, clothes for status-superiority, media fantasy for meaningful work, and so forth. These are all subject to the same analysis of the distinction between desire-satisfaction and pleasure offered above, in Subsection 2.1. I examine the relationship among desire, self-hatred, and consumption, with particular reference to American consumerism, at greater length elsewhere. The important point here is merely that the unlimited proliferation of consumption, like the unlimited proliferation of acts of sadism or revenge, are natural expressions of the two interdependent, highest-order impulses of the Humean self: the desire for wholeness and sufficiency which is impossible to satisfy, and the aversion to self-hatred which is impossible to avoid.

3. Desire and Instrumentality

We have just seen that self-hatred and the consequent impossible desire for status superiority is endemic to the psychology of the Humean self. This self evaluates its perceived external environment with regard to its instrumental opportunities or setbacks, i.e. as indices of power. If the Humean self thus manipulates intrinsic into instrumental goods from the remote but attached perspective of its future satisfactions, then it misuses itself as a similarly instrumental good whose worth is similarly measured in future power obtained. The source of value for the Humean self is determined by the variety and grade of ways in which it can use itself to satisfy its desires: socially, politically, financially.

3.1. The Instrumentalization of Belief

Its own beliefs are, therefore, similarly among the array of instruments and means for satisfying the desires of the Humean self. On the belief-desire model, I begin by having a certain desire. On the basis of my background information and familiarity with the specifics of my situation, as well as my ability to reason and calculate instrumentally, I formulate dispositional and/or occurrent beliefs about how to satisfy it most efficiently. These included beliefs about what my real situation actually is, what resources I have at my disposal, how to utilize them with minimum cost to effect the satisfaction of my desire, and of course, more general beliefs about causal laws, particular causal connections, and so on. Agent-specific means for the satisfaction of desire, then, can be viewed as constituting a two-part, multitiered hierarchy, with theoretical beliefs ranging from the general to the specific in the uppermost half, and the concrete behavior and other resources available to the agent that may express those beliefs in the lower half. Clearly, there is nothing intrinsically instrumental in these beliefs, behavior, or other resources. They become instrumental insofar as the agent places them at the disposal of the desire she wishes to satisfy.

Now assign to the concept of a belief the same degree of ubiquity that Humeans claim for the concept of a desire. That is, think of it as a dispositional or occurrent mental phenomenon that interprets our internal goings-on and thereby causes further ones. Such a phenomenon can take any propositional content whatever, e.g. perceptual (as in "I believe I am seeing a red patch"), emotional (as in "I believe I am angry at having missed the bus"), or intentional (as in "I believe I intend to keep my promise"). This illuminates the sense in which the Humean conception of the self promotes the thoroughgoing instrumentalization of all the constituents of the self. For all such beliefs, on this picture, are available for deployment in the service of the satisfaction of desire, *including beliefs about the intentional object of that desire.* Thus, for example, my desire for security and personal aggrandizement may motivate me more efficiently to surround myself with sycophants, and so to satisfy that desire, if I believe it instead to be a desire for peers who recognize true worth when they see it. Or my desire to inflict pain on a competitor may motivate me more efficiently to exploit his vulnerabilities if I believe it instead to be a desire for excellence in performance at any cost. Or my desire to appropriate ideas from a manuscript I have reviewed, rejected, and refused to return may motivate me more effectively to disregard professional ethics if I believe it instead to be a disinterested desire to improve on the performance of an intellectual inferior. In this way, thoroughgoing self-deception – about my perceptions, emotions, and intentions, as well as about my desires – is rationally justified by the imperative of efficiency inherent in the structure of the Humean self (much as free riding is by the same imperative in the structure of the corresponding Hobbesian society): The rationality of my beliefs, like that of my behavior, is a function of their instrumental efficacy in enabling me to satisfy my desires. No other criteria of rationality are independently relevant, and so no independent moral considerations are, either.

Thus within the constraints of the Humean conception, beliefs may be rational in either of two ways. They may be rational in virtue of representing accurately what I need to do in order to get what I want; call these *veridically* rational beliefs. Or they may be rational in virtue of best enabling me to get what I want; call these *efficaciously* rational beliefs. Veridically rational beliefs are true beliefs about the most efficient strategies for me to adopt in order to satisfy my desires. Efficaciously rational beliefs, on this interpretation, are those which in fact bring about the satisfaction of my desires most efficiently. I may have veridically rational beliefs about what beliefs are efficaciously rational; and I may count veridically rational beliefs among those which are efficaciously rational for me to hold.

Efficaciously rational beliefs may diverge from veridically rational ones because the most efficient action for me to perform in order to satisfy my desires may not be the most efficacious means to the satisfaction of my

desires. It may be that a different set of beliefs, combined with less reflective and deliberate behavior, may be more efficacious in satisfying my desire than acting on true beliefs about the most efficient actions to take to satisfy that desire. It may be more efficaciously rational for me to hold veridically *ir*rational beliefs about the satisfaction of my desires. For example, satisfying my desire to improve the human condition may require false and overly sanguine beliefs about my capacity for satisfying it and about what counts as satisfying it. Or my accurate, detailed, lengthy plan for learning Sanskrit as efficiently as possible may prove to be so boring and pedestrian that it kills my enthusiasm for doing so. The requirements of the Humean conception subordinate veridical rationality to efficacious rationality, because the requirement that I achieve my end as efficiently as possible outweighs the requirement that my beliefs about how to do so be as accurate as possible. Hence in the end, beliefs are truly rational within the Humean conception only to the extent that they are efficaciously rational. Veridically accurate beliefs have no special, noninstrumental value in the belief-desire model of motivation.

This holds not just with respect to beliefs about the various components of the self, but about these components considered independently. Take the emotion of resentment. If I desire to participate more fully in the political process, and find that such activism provides a satisfying outlet for this emotion, which is a pervasive one for me, then it may be rational for me to cultivate and dwell extensively on my feelings of resentment, in order to satisfy my desire to participate politically. In such a case, I use my emotions to motivate me to satisfy a prior, but motivationally ineffective desire. Or take the aural perception of traffic sounds on the street outside my window. If I desire to solve a conceptual problem subliminally, by freeing my imagination from its habitual intellectual constraints, and find that temporary aural distraction enables me to do so, I may attend deliberately to this aural perception in order to solve the conceptual problem subliminally. Here I utilize a perception to satisfy a desire I would be unable to satisfy by attending to the object of the desire directly. As we have already seen, desires themselves are equally susceptible to this brand of instrumentalization in the service of a further desire, as when I recruit my long-term desire to master a chunk of philosophical material in order to satisfy my immediate desire to finish preparing a lecture.

These cases, and others like them, are not unusual because they are unfamiliar or infrequent, but rather because they are rationally *prescribed* by a model of motivation that stipulates the satisfaction of desire as the only motivationally effective source of intentional behavior, and that behavior itself – any behavior – as reflecting the agent's beliefs about how best to go about this. It is rationally prescribed because, as we have seen in Subsection 2.1, above, no instrumental resources, whether internal or external to the

physical boundaries of the agent, are exempt or privileged with respect to this stipulation. *The desire model of motivation requires the mobilization of all the components of the self - physical and mental, internal and external – as potential resources for the appropriation of objects of desire.* In Chapter IV.5 I argue that the Humean model of rationality is incapable of accommodating moral or rational side-constraints as intentional objects of behavior. This means that these other components, including, for example, a sense of duty or moral conviction, feelings of compassion or moral indignation, and so on, are not just instrumental to the satisfaction of desire, but motivationally subordinate in importance to it as well. Call this the *instrumentalization dilemma.*

3.2. The Instrumentalization Dilemma

This dilemma is exemplified by the case of Dick. Suppose Dick desires to become a spontaneous and emotionally responsive person, adept at discerning his emotional reactions and at articulating them honestly to his friends. Suppose further that this desire is instrumental to a further desire to improve morally his personal and social relationships by being concerned, compassionate, and honest. To these ends, Dick undergoes therapy, keeps a journal, and encourages his friends and associates to confront him with their responses to his behavior, and to engage with him actively regarding whatever issues are raised by doing so. He realizes he is inviting intense emotional upheaval by seeking out such situations and analyzing them introspectively. But his desire to improve morally his interpersonal relationships is genuine, and he strongly believes (correctly, let us suppose) that his emotional arridity and fear of vulnerability have made him a moral cripple in the past. Then suppose Dick reacts dismissively or arrogantly in a meeting to a professional associate's suggestion as to how to improve the efficiency of their business, and is taken to task for it publicly. It is suggested that Dick frequently has difficulties in countenancing from women colleagues the same professional input that he invites from men; is even more characteristically ungenerous when competing with the former; that perhaps he feels threatened by women, or has not successfully resolved his separation from his mother, or is re-enacting his childhood sibling rivalry, and so on. Dick's responses to these confrontatory remarks are various: He is alternately outraged, thoughtful, defensive, receptive, insulted, and sarcastic. Occasionally he is sorely tempted to storm out of the meeting in a huff, or revamp his strategy for moral self-improvement; but is reminded, or reminds himself, of his commitment to the process of social engagement – and so sits it out, outwardly contemptuous of his colleagues' unconscionable armchair psychologizing, but inwardly wondering whether they may not, after all, be right.

Now most of us know such individuals, and it is worth examining why we may feel an uneasy sense of insincerity in their presence. This response

itself is a complex one: We ourselves may feel insincere for not revealing our own personal foibles and blind spots more fully, in the presence of someone who courageously subjects his imperfections and immaturities to the traumatic ordeal of public scrutiny, the way the matador waves the red cape before the bull. We may also feel that there is something distastefully exhibitionistic in this public display of breast-beating, and that suffering one's neurosis silently is more honorable. But there is often more to the response than this. We may, in addition, sense something insincere in Dick's own stance toward his psychological and moral flaws: If they are all out there on the table, then who is in the kitchen? That is: what *kind* of psychological entity is serving them up?

The problem is that despite Dick's avowals, the belief-desire model of motivation requires us to view all his overt behavior, including his actions, avowals, responses, and explanations, as instrumental to the satisfaction of some further, unspoken desire. So by that hypothesis, none of that behavior intrinsically expresses the desires Dick *says* he is striving to satisfy. Dick's stated program for moral self-improvement requires the assumption that he has, as it were, turned himself inside out for the sake of that program; that the reactions we and he are invited to scrutinize are not just instrumental to the satisfactions of the self, but expressive of it. But by the lights of the belief-desire model of motivation, we are entitled to view this assumption with suspicion. Because no matter how genuine and transparent his reactions, they are mere instruments to the satisfaction of some further desire. And so Dick's motivations for revealing them remain opaque: they are, by hypothesis, not among the responses with which we are invited to engage.

Of course Dick explains his moral reason for inviting us to engage with him in this way. This explanation, too, is among the issues raised by his behavior with which we are asked to engage. But we have no independent evidence for the truth of this explanation, and no reason to accept it on faith, as he seems to want. Instead it merely defers our suspicions one remove, rather than allaying them. For now the question becomes that of what desire he satisfies by adopting this strategy for moral self-improvement rather than some other; what desire he satisfies by telling us all this; and to the satisfaction of what desire Dick's desire for moral self-improvement itself might be instrumental. We may think him excessively self-absorbed, preoccupied, narcissistic, or simply a glutton for attention, thereby discounting right away the possibility that his concerns are authentic. Indeed, the more confiding and self-revelatory Dick becomes, in response to our demurrals, the more we may feel somehow sucked in or manipulated; and the correspondingly less of a chance he has to satisfy his stated desire to improve his moral condition and relationships with others.

So far the instrumentalization dilemma has been painted as arising from our third-personal perspective on Dick's avowals; from an apprehension that

he is in fact instrumentalizing his relationships with us to serve an agenda to which we are not privy. But we can see that Dick himself may not be impervious to these concerns. He may wonder, for example, whether his stated moral object of desire is what is really structuring his responses and behavior instrumentally, or whether it may not in fact be merely a valued side-effect of his narcissism, or his desire for attention or power. For given the premise that it is an object of desire to which all the other components of his self are instrumental, it remains an open question which of the available, conjoined ones is really doing the structural and motivational work; or whether it is among the available, conjoined ones that he should even seek an answer.

But Dick's deeper concern is an unease at the very availability of his thoughts and responses for instrumentally rational scrutiny. Here the worry is not just the obscurity of the overriding object of desire as such; but rather that the explicit desire from the perspective of which he introspects, whatever it is, preselects which thoughts and responses are available for such scrutiny, and thereby obscures the yet more basic intrinsic desires that are more thoroughly in need of reform. The distance afforded by the intrinsic object of desire to which Dick is apparently committed makes the true explanation of his behavior seem equally remote and inaccessible. His unease is a consequence of the truisms into which the desire model of motivation inevitably degenerates, namely: You can only know about yourself what you *desire* to know; and: What you desire to know about yourself is never what you truly *need* to know. Dick's worry that he may be overlooking the elements in his personality that are truly responsible for his moral and social sterility are a natural consequence of his suspicion that it is precisely the self he desires to reform that is engendering that very desire instrumentally, as a kind of lip-service to moral self-improvement that makes genuine improvement impossible. Thus our suspicion of his insincerity, and sense that we are being manipulated, may be mirrored in Dick's own worries that his true motives remain obscure, and his character beyond the reach of his own conscious efforts at self-development.

3.3. The Instrumentalization of the Self

Note, then, that the instrumentalization dilemma is not generated by assuming the nonmateriality or privacy of Dick's desire for moral self-improvement, or whatever desires are hypothesized to motivate that one. The dilemma would remain even if we could literally see Dick's internal states represented on a video screen as states of his organs and brain. For the question would still arise as to how these states should be interpreted, and what explanatory hypothesis was best suited to this purpose. The hypothesis in question, namely the belief-desire model of motivation, stipulates that all such states are resources for the satisfaction of desire, and it has already been

argued that this must include available desiring-states as well. This means that any such state that appeared on the video-screen could be assumed to be instrumental to the satisfaction of some further desiring-state, which therefore, by hypothesis, did not appear on the video screen; and the desire model of motivation encourages us to make this assumption.

Nor would the instrumentalization dilemma be solved by assuming the existence of some ultimate desiring-state – such as that for wholeness and sufficiency – that did not appear on the video screen, but rather somewhere else, or only to Dick. For wherever, or to whomever it appears, the question can always be reiterated: To the satisfaction of what further desire is this one instrumental? Dick's underlying motives in doing what he does are obscure, not because they are private or nonmaterial, but rather because they are by hypothesis one step ahead of whatever motivational explanation we already have. This lays the foundation in the arena of psychological explanation for Chapter VIII.2's analysis of the infinite regress problem of self-evaluation that many have noted besets the Humean conception of the self. The existence of a highest-order terminating criterion of rationality for the Humean conception – i.e. the desire for wholeness and sufficiency – terminates the regress of orders of desire. But it does nothing to terminate the regress of instrumental desires themselves. There is, then, no point at which we can rest assured that all of his motives in invoking our participation are fully "on the table." We are entitled to probe just as much further when he represents them to us on the video screen as when he explains them to us verbally.

Postulating a Freudian unconscious is equally unable to allay Dick's and our worries, although it is a natural consequence of accepting the desire model of motivation to turn to the Freudian variant. The premise that all mental and physical behavior is instrumental to the satisfaction of a further, intrinsic desire obfuscates both the object of that desire and its motivationally effective desiring-state simultaneously. Then it may indeed seem that if these motives and objects are not accessible to our scrutiny, they must exist somewhere else – the unconscious – where no one, not even the agent, can get at them. But this does not nullify the concern that motivated this tack, for that a desire exists in an epistemically inaccessible psychological realm does not imply that it is therefore not instrumental to the satisfaction of some further, equally inaccessible one. So it is possible to view one of the central postulates of Freudian theory as a natural, though not conceptually inevitable, consequence of an historically and conceptually prior commitment to the Humean conception of the self. For it is not useful, in order to satisfy the explanatory requirements of the belief-desire model, merely to stipulate that final desires which are inaccessible to us here exist somewhere else, where they *are* inaccessible. That they are inaccessible *there* does not alter their inherently instrumental designation.

The instrumentalization dilemma is generated, then, not by Dick's moral interests, nor even by his strategy for achieving them. For if we and he were inclined to take his words and deeds at face value, we could each simply decide whether or not it was a good project, whether he should go through with it, and whether we wanted to participate or not. Rather, Dick's difficulties are engendered by his – and our – presupposition of the belief-desire model of motivation in formulating and assessing it. By conceptually nesting his behavior, and his and others' responses to that behavior, within a scheme relative to which all such events are instrumental to the satisfaction of an ultimate desire he is assumed to have, we collaborate with him in effectively obviating the possibility that that desire will be satisfied; or, if it is, that anyone will be able to recognize it as being so. For the self that desires that satisfaction itself remains, by hypothesis, impervious to the transforming effects of our common scrutiny.

And so it must, again by hypothesis. For whatever occurs, either in Dick himself or in us, are at best instrumental or constitutive means by which his ultimate desire is satisfied; they are not themselves direct expressions of that desiring. Thus yet another infinite regress arises in response to the express desire for self-knowledge or self-assessment, which is entirely familiar, predictable, and self-defeating: For any proposition *P* that I (or you) may entertain about my self as true, it is equally true that perhaps I (or you) believe *P* only because of my (or your) desire *Q*. But my (or your) belief in my (or your) desire *Q* may be explained entirely by its efficacy in satisfying my (or your) further desire *R*. And so on. This schema may deny us the Olympian satisfaction of irrefutable self-knowledge; but it simultaneously affords us an endless series of "perspectives" and "insights" on our actual behavior, with which we may entertain ourselves endlessly, from the perspective of that hypothesized desire we currently acknowledge as final.

In actual fact, we need not pursue the regress doggedly, in order to arrive at a satisfactory explanation of a person's behavior; and, unless we are feeling particularly perverse or powerless, we usually don't. Rather, we accept that explanation that best coheres with our other beliefs about her, and call it into question only with the acquisition of further beliefs with which it may fail to cohere. In these commonsense cases, the explanation in question need not invoke a "deep" desire or other motive. An agent's absentmindedness, insensitivity, or naiveté often suffices to explain behavior that the desire model of motivation encourages us unendingly to probe. In Volume II, I develop an alternative model of motivation that tries to better respect these ordinary psychological facts about us. For now it should be noted just that the belief-desire model of motivation as stated encourages us to regard any attitude or desire we currently ascribe to the agent as instrumental to a further one. Thus our suspicion of Dick's insincerity, and that we are being manipulated by his stated moral program, are built into the Humean

conception of the self. He has framed his project in such a way that, given this set of metapsychological assumptions about the self, we are all forced to the conclusion that he is hiding something; and using his and our responses in the service of that which he is hiding.

3.4. The Puppeteer Fallacy

Of course this conclusion is mistaken. It relies on a suspect view of the "real" self as puppeteer, pulling our psychological strings as subjects, agents, and observers from behind the scenes, for some further purpose to which we are, by hypothesis, not privy. Call this the *puppeteer fallacy.* This fallacy does not arise because of the supposed nonmaterial status of internal states, as Ryle thought. As we have already seen, the identification of motivationally effective desires is equally obscured by the stipulation that desires are physical events to which all physical behavior is instrumental. Nor is the problem engendered by the purported privacy of desire, as Brandt and Kim's dispositional analysis of desire as a theoretical construct without experiential analogues seems to suggest. We have also seen that a person's public behavior and declarations are equally as inscrutable, on this view, even to the agent, because equally instrumental to the satisfaction of a desire that is not just physically or publicly but *conceptually* obscure.

From this perspective, a major complaint against the belief-desire model of motivation is that it makes us out as agents to be much more in control of our behavior than we are, and thus engenders the insoluble puzzles about free will and determinism I address in Chapter VIII.2. By promoting the ascription to the agent of a hidden agenda to which all of its – and our – mental and physical behavior is instrumental, it exacerbates the puppeteer fallacy by rendering the "true" self simultaneously ubiquitous and elusive. And this confronts us all with the triply self-defeating (literally) task of discerning what that hidden agenda is, whether or not we approve of it, and to what extent it should indeed be promoted or discouraged.

This model thus indirectly instrumentalizes all social relations. For in response to this perceived hidden agenda, it forces us to construct our own. Acceptance of this model encourages us to regard unfamiliar individuals as cryptic or unpredictable, and familiar ones as secretive, devious, or manipulative; and we ourselves often choose our words and gestures with an eye to their calculated psychological or social consequences, rather than their conventional linguistic meanings. For of course the supposition of hidden desires that motivate an agent's social behavior requires us as participants to respond to those supposed desires, and not to the overt social behavior that is interpreted as instrumental to them. It thereby requires us to choose our responses on the basis of calculations of how best to reinforce, modify, or discourage those desires in light of our own.

This shared imperative may even generate a second-order set of meta-conventions of social meaning, supervenient on the traditional ones, in which linguistic participants attempt to arrive at a mutual understanding of one another's postulated hidden desires, by inference from utterances and behavior which, to the uninitiated, are completely unrelated to the desires with which they are, according to the meta-conventions, systematically correlated.[27] Of course the development of meta-conventions of social meaning serve merely to up the ante in terms of the efficaciously rational calculations each agent must perform in order both to comprehend and communicate successfully, and the degree of efficaciously rational sophistication required in order to manipulate these meta-conventions in the service of one's own desires. For on this model, there is invariably some motivational variable that must remain unspoken. But it is not difficult to imagine actual situations of bargaining, diplomacy, or social coordination in which repeated exposure and practice may instill these skills of perception, calculation, and response so deeply in one as to be second nature.

Thus the impersonal attitude that Strawson[28] argued to be a consequence of taking determinism seriously as a theory of human social and moral behavior can be gotten by a much shorter route. Regardless of whether or not human behavior and motivation is causally determined, a commitment to the belief-desire model of motivation as an explanatory theory has the same outcome; for as we see in Chapter XII below, it replaces the ideal of social cooperation with strategies of mutual manipulation. It encourages us to disregard the prima facie significance of what others actually say and do, in order to seek out their underlying desires and shape them for our own ends. For any direct appeal to their rational or moral faculties is presumed to be itself instrumental to the satisfaction of some further desire.

The unreflective acceptance of this model makes it unsurprising that we approach the practical tasks of self-knowledge and self-control with the aid of a professionally trained, paid third-personal perspective – whether therapist, self-help manual, or religious counselor, whose job it is to infer from our mental and physical behavior the true, hidden condition of the self, and to prescribe remedies for healing it. In each of these cases, we presuppose the ubiquitous existence of hidden desires the identification of which provides the key to mental and physical behavior hypothesized to be instrumentally rational to them. These desires are hidden to the observer because we perceive only the agent's physical, instrumental behavior. They are hidden to the agent

[27]For accounts of meaning that might be compatible with this analysis, see H. P. Grice, "Meaning," *Philosophical Review 66* (1957): 377-88; and Stephen Schiffer, *Meaning* (Oxford: Oxford University Press, 1972).

[28]P. F. Strawson, "Freedom and Resentment," in *Freedom and Resentment and Other Essays* (London: Methuen and Co., 1974).

because, on the Humean view, even our mental behavior as agents is instrumental to the satisfaction of final and ultimate desires to which we by definition lack access. Thus for both, this presupposition requires a detached, third-person perspective on those desires that is personally invested neither in their frustration nor their satisfaction.[29] The remote but attached perspective of future wholeness and sufficiency from which the Humean self regards its present condition of want is not adequate to fulfill this requirement. Under these circumstances, verifiable self-knowledge becomes a theoretical impossibility, the act of trust required for genuine friendship a fundamentally irrational leap of faith, and moral concern and personal honesty objects of desire that seem remote indeed.

4. The Veracity of the Model

In this chapter I have tried first to sort out and clarify the conception of desire on which the belief-desire model of motivation – and so the Humean conception of the self – in fact rests. I have argued for the inadequacy of both the orthodox and the revisionist variants on this model, particularly as they find expression in the work of Brandt and Kim, Goldman, and Lewis; and have proposed to replace them with a representational analysis of desire that circumvents their liabilities yet retains their assets. I have then tried to explicate some of the psychological, characterological, and behavioral implications of the representational analysis for actual selves socialized and structured in accordance with the Humean conception. I have concluded that a Humean self is committed to the satisfaction of several Quixotic and structurally impossible higher-order desires: for wholeness and sufficiency, for the avoidance of self-hatred, for status, power, recompense, unlimited consumption, and self-knowledge.

Of course the fact that these objects of higher-order desire are impossible to attain does not imply that actual human agents do not desire them nevertheless. And so it may be objected to this critique of the belief-desire model of motivation that the frequency with which these implications are confirmed in actual human behavior within a global consumerist culture redounds to the credit of the Humean conception as a plausible and well-confirmed explanatory hypothesis, rather than undermines its legitimacy as the critique apparently intends. But recall from the General Introduction to this project that this critique is embedded in a more general one that argues not that the Humean conception of the self is intrinsically wrong, but rather that it is incomplete; that it is inadequate to the *full range* of psychological facts of human nature, and so often makes false predictions about human behavior; and that its internal, structural and conceptual inconsistencies arise from the

[29]See Peter Alexander "Rational Behavior and Psychoanalytic Explanation," in Care and Landesman, *op. cit.* Note 3.

attempt to universalize an explanatory hypothesis that in fact is incapable of universalization. So the plethora of empirical instances that confirm the psychological profile I have developed here do not undermine the critique unless no counter-instances can be found to disconfirm it. And I have tried in Section 1, above, to demonstrate that this question cannot be settled simply by arbitrarily extending the denotational scope of the term "desire" to include all human motivation. Humpty Dumpty couldn't get away with it, and neither can we.

Chapter III. The Utility-Maximizing Model of Rationality: Informal Interpretations

In Chapter II I examined the motivational model of the Humean conception of the self, the belief-desire model. In this and the following chapter, I examine its structural model, the utility-maximizing model of rationality, and try to make good on the promissory note issued in Chapter II. There I suggested that the revisionist view of desire as a theoretically ubiquitous explanatory entity rendered it vacuous – and, along with it, the utility-maximizing model of rationality within which desire is embedded as its sole conative element. Here and in Chapter IV I contend that in order to retain its status as a *bona fide* explanatory theory, utility theory also must abdicate its claim to universality. This conclusion follows, I claim, upon the Humean refusal to impose substantive constraints upon the final ends – or intrinsic preferences – the attainment of which constitute maximizing utility. Without such constraints, any behavior can be interpreted as utility-maximizing because any final end can be understood as a source of utility, and hence any behavior can be rationalized through the ascription of such a final end to it. This makes the theory universal in its explanatory reach, all right, but also vacuous. Only by imposing such constraints can some ends – and therefore some behavior – be identified as irrational in its terms, i.e. not susceptible to interpretation as a case of maximizing utility. These constraints thus protect the theory from vacuity, but only by sacrificing its claim to universality.

Some economists would question the need to demonstrate this. They take it to be obvious that this model of rationality is intended to ground a specifically economic theory of consumer behavior under free market conditions. They take it to be equally obvious that human beings do not act as free-market commodity consumers in all areas of their lives; and that the utility-maximization model therefore has a restricted scope of application. I agree with this view. But we have already seen in Chapter I how philosophy generally tends to seek outside the boundaries of its own discipline for scientific models that can be imported back into it and pressed into foundational service. In this regard, metaethics – particularly Humean metaethics – is no different. In subsequent chapters I scrutinize Nagel's, Gewirth's, Rawls's and Brandt's success at this; but they are only a few of the contemporary metaethicists who look to the Humean model of rationality for a scientifically validated foundation on which to erect a well-justified normative moral theory.[1] So in the present and the next chapter, I develop at

[1] Among the many late twentieth century moral philosophers who have promoted this view are Robert Nozick, *Anarchy, State and Utopia* (New York: Basic Books, 1974); and

greater length the thesis that the universalized version of the utility-maximizing model of rationality is vacuous, by addressing several different interpretations of the Humean claim that utility theory has explanatory universality. For the most part, my criticisms of this claim address the purported scope of the model, not its substance; and therefore have no bearing on the model's validity within the context of free market economics for which, in its formalized version, it is intended.

Similarly, I do not try to specify which final ends should be excluded from the scope of a suitably restricted version of the utility-maximization model of rationality; nor, therefore, which areas of human action may lay entirely outside it. So my conclusions here do not imply that the utility-maximizing model may not have important applications to such areas as making financial investments, purchasing lottery tickets, or buying insurance. My argument is merely that the utility-maximizing model of rationality must incorporate some such restrictions, in order to avoid vacuity. However, the argument does, therefore, imply that Max U. (as some economists fondly refer to the quintessential utility-maximizer) cannot do just *anything* to maximize utility. So if you agree that, for example, Max U. should not be allowed to harvest body parts from the poor at 50¢ a piece and sell them at $50,000.00 to the rich for organ transplants, you should have some sympathy for my thesis.

Whereas Chapter IV focuses on formal decision-theoretic interpretations of the utility-maximizing model, the present chapter focuses primarily on its informal philosophical arguments. Section 1 specifies the formulation of the basic principle of utility-maximization (U) that I target both here and in Chapter IV. Sections 2 and 3 develop in detail the criticism that if (U) is formulated so as to have universal application as an explanatory social theory, then it is either vacuous, in that it is confirmed by all instances of behavior; or else its detailed formulation is internally inconsistent, in that it implies that utility-maximization itself both is and is not an intentional end, subject to cost-benefit analysis like any other. The argument begins with the simplest and most commonsense rendering of (U); and proceeds by examining increasingly complex and sophisticated formulations of it. I show in Section 2 that this

---

David Gauthier, "The Social Contract as Ideology," *Philosophy and Public Affairs 6* (1977), 130-164. Some philosophically inclined jurisprudes are also guilty. Mid- to late twentieth century work in the economic analysis of law applied the utility-maximizing model of rationality to the legal sphere. See Ronald Coase, "The Problem of Social Cost," *Journal of Law and Economics 3* (1960); "Durability and Monopoly," *Journal of Law and Economics 15* (1972); and Richard Posner, *The Economic Analysis of Law* (New York: Little, Brown, and Co., 1975). For a critical view, see Staffan B. Linder, *The Harried Leisure Class* (New York: Columbia University Press, 1970); Tibor Scitovsky, *The Joyless Economy* (New York: Oxford University Press, 1977); and Robert Paul Wolff, "Robert Nozick's Derivation of the Minimal State," in Jeffrey Paul, Ed. *Reading Nozick* (Totowa, NJ: Rowman and Allenheld, 1981), 77-104.

implication holds for single ends considered independently; and in Section 3 that it holds equally for the coherence set comprising all of one's ends. In Section 4 I show that this argument holds whether the concept of utility is interpreted phenomenologically, psychoanalytically, or behaviorally. The following chapter then extends these conclusions to more technical formulations of (U).

1. Formulating the Principle

My target is one particular formulation of the basic principle of utility theory, namely

(U) If a rational agent acts, she maximizes utility.

This minimalist formulation of (U) superficially resembles what Maurice Allais calls the Axiom of Absolute Preference, i.e. that given two alternatives, a rational agent prefers that one that consistently yields the greater gain.[2] However, Allais claims that this axiom is merely a consequence of a much weaker definition of rational choice as consistency. But the three criteria that constitute his definition always include the axiom of absolute preference among them.[3] Without it, there would be no ordering relation specified for the ordered set of gambles among which, on his definition, a rational agent chooses. I criticize his conception of consistency directly below. Moreover, Allais' Axiom of Absolute Preference is not, according to his account of it, implied by the assumptions grounding the von Neumann-Morgenstern cardinal utility function (discussed in below in Chapter IV, Sections 1.2-3). By contrast, (U) is presupposed by it, in the concept of a highest-ranked member of an ordered set of preference alternatives, regardless of whether these alternatives are objects, events, conditions, states, or gambles.

(U) is similar to Allais' Axiom of Absolute Preference, however, in its atomistic ascription of utility-maximization to the smallest behavioral unit in which preference is revealed. It is also similar in assuming satisfaction of the

---

[2] See his "Fondements d'une Théorie Positive des Choix Comportant un Risque et Critique des Postulats et Axiomes de L'Ecole Americaine,"Memoir III of *Econometrie XL* (1953), 257-332 (Colloques Internationaux du Centre National de la Recherche Scientifique, Paris), translated as "Foundations of a Positive Theory of Choice Involving Risk and a Criticism of the Postulates and Axioms of the American School," in Maurice Allais and Ole Hagen, Eds. *Expected Utility and the Allais Paradox* (Dordrecht, Holland: D. Reidel, 1979), 27-146. See esp. 39-41.

This formulation of (U) also encapsulates Ward Edwards' treatment in his comprehensive survey paper, "The Theory of Decision-Making," *Psychological Bulletin 51*, 4 (1954), esp. 381-3.

[3] Allais, *ibid*. See particularly 34, 69, 78-79, 82, and footnote 78'.

Independence of Irrelevant Alternatives condition on preference orderings, what McClennen would call the Context-Free Ordering.[4] This condition states that an ordering Y produced by a series of pairwise comparisons among a set of alternatives X is not changed by the introduction of additional alternatives on some particular occasion that expands X to a larger set X*. So, for example, if I prefer alternatives P to Q, Q to R, and R to S, the introduction of alternative T, on a particular occasion, does not alter my ranking of P, Q, R, and S relative to one another. Either I prefer T to P, or I prefer S to T, or I do not rank T at all. Thus this condition excludes the case in which the introduction of T leads me to reverse my ranking of Q and R. This means that the ranking of a set of alternatives can be inferred from its pairwise rankings and vice versa. The discussions below, in Chapter IV, and in Volume II, Chapter III follow Sen[5] and Broome[6] in assuming that considerations of temporal continuity and human cognitive limitations necessitate pairwise comparisons as a necessary precondition of Y, and thereby block the inference to the material equivalence of Y and the series of pairwise rankings that produce it. Nevertheless, each can be read off from the other.

(U) is also superficially similar to a different and more nuanced formulation, that if a rational agent acts, she maximizes her preferences, that would seem to avoid the vacuity I argue to be endemic to the utility-maximization model. On this conception, the agent's preference rankings are numerically represented by the theorems of utility theory interpreted as representation theorems in a theory of measurement, and she chooses from among those alternatives she maximally prefers. The ordering axioms – transitivity, connectedness, asymmetry of strict preference – then ensure the existence of at least one most-preferred alternative in a finite set of such alternatives, and the agent maximizes by choosing that alternative. As Sen has shown that connectedness is not necessary for maximizing in this sense, it can be replaced by his concept of a maximal set, i.e. sets consisting of mutually nondominating alternatives none of which is strictly dispreferred to any other.[7] The transitivity requirement similarly can be replaced by a weaker requirement of acyclicity of strict preference, since maximization requires merely the avoidance of cyclical preferences. It would seem, on the face of it,

---

[4] Edward McClennen, *Rationality and Dynamic Choice* (New York: Cambridge University Press, 1990), xi, 29-31, 64-67.

[5] Amartya K. Sen, *Collective Choice and Social Welfare* (San Francisco: Holden-Day, Inc., 1970), 3.

[6] Broome, John, "Rationality and the Sure-Thing Principle," in *Thoughtful Economic Man*, edited by Gay Meeks, Cambridge University Press, 1991, pp. 74–102. Cited and discussed in McClennen, *op. cit.*, 66-67.

[7] Sen, *op. cit.* Chapter 1*, Section 1*2.

that this conception is weak enough to satisfy universality yet strong enough to avoid vacuity.

However, a theory of measurement adequate to this conception would first have to solve the problems raised by interpersonal comparisons that are discussed in Chapter IV, Sections 1.1 – 1.4 below. Second, weakening transitivity to acyclicity does not alter the behavioral options actually available to the choosing agent, because the two are logically equivalent; though a demonstration of this must await the apparatus I develop in Volume II, Chapter III.6.2.1. Third, the argument of Chapter IV, Section 1.6 below implies that it is in any case not possible to exclude cyclical preferences by imposing *any* further familiar normative requirements – neither transitivity, nor irreflexivity, nor independence, nor substitutability, nor continuity – unless these are subordinated to strictly logical constraints on preference orderings for which the canonical notation of decision theory affords no resources. I offer some in Volume II, Chapter III. Finally, a closer look at the money pump in Chapter IV, Section 2.3 shows that utility-maximization does not require excluding cyclical preferences in the first place. So the criticisms of (U) I make in the following pages apply to this more complex variant on it as well.

(U) is not the only possible formulation of the principle. Some utility theorists would insist, on Bayesian grounds, that (U) should address the maximization of expected utility. Whether (U) is formulated so as to address the maximization of utility or of expected utility does not, for the most part, affect the substance of my arguments. Unless explicitly stated otherwise, my arguments do not depend on whether a rational agent is assumed to act under conditions of certainty, risk, or uncertainty. Rather, they focus on the concept of a preference ordering that must be presupposed in the assignment of objective as well as subjective probabilities to options. But I shall assume for the sake of argument that the agent has full information and that any probability assignments to outcomes are based on multiple trial repetitions, again unless specifically indicated otherwise.

Others would substitute "human beings" for "a rational being," emphasizing the descriptive over the normative and reducing the explanatory scope of (U) in that way. Still others would object to the suggestion of intentionality in (U), on the grounds that action implies intentionality and agents maximize utility whether or not they intend to. I think it is a mistake for a Humean to raise this objection, for reasons explicated in Section 4 below. Yet others would complain that (U) conceals an essential normative dimension, in that people should maximize utility but often do not in fact. I address expected utility theory's conception of the relation between what human beings actually do and what fully rational beings are conceived to do in Chapter IV, Section 4, below; and the issue of intentionality in this chapter's Section 2, below.

The point of choosing this particular formulation of the basic principle is to grant the utility theorist as much as possible at the outset. Among the philosophical claims that are often made for utility theory are: that it enables us to predict actual human behavior; that it furnishes a systematic account of fully rational behavior; and that it supplies a normative guide to rational choice to which actual human behavior aspires and may approximate to varying degrees. To the extent that actual human behavior achieves the status of rational behavior, then, an axiomatized system of decision theory should enable us to predict it. (U) captures the idea of utility-maximization as fully intentional and therefore deliberate behavior. Furthermore, it captures the idea of a basic action (in Danto's sense) – specifically, the type of basic action involved in making pairwise comparisons among a given set of alternatives – as itself utility-maximizing prior to any more complex utility function that may emerge from such an ordering.

(U) is thus charitable both to our actual human potential for rational behavior, and to the potential of axiomatized systems of decision theory to predict such behavior. I argue, however, that if (U) is assumed to be universal in its scope of application, then regardless of the further details of its interpretation, it is either a tautology of the form P → P, or else logically inconsistent. If it is either then it cannot meet the requirements of an explanatory theory because it "explains" everything and nothing simultaneously. But if (U) is reconceptualized as a contingent principle of limited scope,[8] subordinate in status to the requirements of logical consistency, then any reformulation of (U) needs to observe the constraint of logical consistency as well.

---

[8] Harvey Liebenstein reconceptualizes (U) in a very different direction in his *Beyond Economic Man* (Cambridge, Mass.: Harvard University Press, 1976). His attempt to reformulate microeconomic theory in terms of a basic concept of trying or "effort" to replace that of maximization avoids some of the difficulties I discuss below – as H. A. Simon's concept of "satisficing" does not (see his "A Behavioral Model of Rational Choice," *Quarterly Journal of Economics 69* (1955), 99-118; and "Rational Choice and the Structure of the Environment," *Psychological Review 63,* 2 (1956), 129-38, discussed in Note 20 below). In *Hard Choices* (New York: Cambridge University Press, 1986), Isaac Levi subverts (U)'s claim to universality by demonstrating the inadequacy of revealed preference theory to characterize decision-making under the very widespread condition of unresolved conflict: Since its favored conception of optimality does not require a weak ordering of all the elements of the choice set, it recognizes optimal alternatives even when that set contains ordinal conflict among some of them. Hence revealed preferences may generate alternatives that are optimal without being admissable in a given value-structure. I myself mount complaints about the revealed preference interpretation of (U) in Chapter IV, Sections 1 and 2, below. Much less than being, in I. M. D. Little's words, "nothing other than a purely logical exercise" ("A Reformulation of the Theory of Consumer's Behavior," *Oxford Economic Papers I* (1949), 99), universalized utility theory, according to the main thesis of this chapter, is *not even* a logical exercise.

This makes (U) and its alternative reformulations a special case of a more general, Kantian theory of rationality that situates the principle of noncontradiction at its foundation. However, by this I do not mean what Allais means. For him, action that is "non-self-contradictory" and so rational satisfies two criteria: first, its ends are "logically consistent;" and second, its means are appropriate to them. "Logically consistent ends" for Allais are those which constitute a set ordered by his axiom of absolute preference, i.e. such that its elements are ordered by the relation "<<".[9]

I agree with Allais' insistence that consistency is the only criterion for the rationality of ends, which are otherwise arbitrary, and with his sensitivity to legislating any more substantive, "politically correct" criteria for the rationality of ends.[10] But his account would not suffice for logical consistency as philosophers ordinarily use that term; nor would the more general account of non-self-contradiction in which his notion of logical consistency figures. In the philosophical context, two sentences are consistent if and only if one does not contradict the other. We ascertain this by quantifying them using the conventions of predicate logic, and relating them by means of the traditional Boolean connectives ".", "v", "→", and "~". However, the ends that constitute elements in Allais' ordered set are not quantifiable using the symbolic resources of predicate logic, and "<<" is not one of the Boolean connectives. Below in Chapter IV, Sections 2 and 3, I argue that no notion of consistent choice that does not meet these two basic philosophical desiderata can do the job even on its own turf; and in Volume II, Chapters II and III of this project I try to develop one that does.

2. The Single End Interpretation of (U)

I describe as the *single end* interpretation of (U) the commonsense notion that a rational agent maximizes utility if he acts efficiently to achieve a particular goal, i.e. by minimizing the expenditure of resources in its service (this is efficiency in the pedestrian rather than the Pareto sense). Can any particular action fail to achieve its particular end efficiently in this sense? I conclude that either no action can, in which case (U) is vacuous; or else the concept of efficiency is inconsistent.

In the single end interpretation of (U), we rely on an implicit *ceteris paribus* clause, by evaluating the rationality of an action in the service of one particular end, assuming all others to be fixed. Conventionally, this interpretation finds expression in questions as to whether a particular action is the most *efficient* way to achieve a given end. According to the pedestrian version of the concept of efficiency, we achieve such an end efficiently when

[9] Allais, *op. cit.* Note 2, 40.

[10] *Ibid.*, 70 and footnote 52.

we minimize the expenditure of resources in its service, irrespective of other substantive ends we may have. The requirement to do this can be understood as following naturally from two weak assumptions: (1) that resources are limited; and (2) we have other ends. But even in a case in which both of these assumptions were false, i.e. in which an agent had just one end – for example, making money – and more than sufficient resources for achieving it, he would still have reason to minimize their expenditure. For the one end he had might still require adopting and achieving subsidiary or instrumental ends in order to achieve that one. Squandering resources on one part of his overall action plan might well leave him unable or less well equipped to bring about the others. More than sufficient resources are not the same as infinite resources, so he would still have to trade off their expenditure on some of his instrumental ends against their availability for achieving others. So minimizing expenditures in order to maximize achievement would be rational whether (1) and (2) held or not.[11] In either case, this concept of efficiency is roughly interchangeable with that of utility-maximization.[12]

For example, dissecting the gluttonous spending patterns of the United States military by contrast with the deeply acculturated thrift of its declining manufacturing industry, Jane Jacobs remarks about industrial engineers that they are "major antagonists of waste and inefficiency … [t]heir objects are to maximize efficiency and minimize costs."[13] Similarly, the Waterford Crystal Company claims (unpersuasively) to *repudiate* efficiency in this sense by announcing that

> At Waterford, we take 1,120 times longer than necessary to create a glass. While a machine can churn one out in only 45 seconds, we take over 14

---

[11] For this formulation of the argument, and at many points in this chapter I am grateful to Ned McClennen (Personal Communication, July 9, 1991), and to his *Rationality and Dynamic Choice: Foundational Explorations* (New York: Cambridge University Press, 1990), from which I have learned a great deal. I discuss the significance of McClennen's concept of resolute choice in Volume II, Chapter IV.

[12] *Pace* Harsanyi, who complains that "the means-end concept of rational behavior is too narrow because ... it restricts rational behavior to a choice among alternative *means* to a given end, and fails to include a rational choice among alternative *ends*" ("Advances in Understanding Rational Behavior," in John Harsanyi, *Essays on Ethics, Social Behavior, and Scientific Explanation* (Dordrecht: D. Reidel, 1976), 93. Harsanyi's criticism seems to ignore the ontological insecurity of the distinction between means and ends: Means or resources to achieve our final ends are themselves instrumental ends, and the mutual adjustment of final ends so as to preserve coherence and ensure their achievement is itself a means of maximizing utility. The discussion of opportunity costs that follows Harsanyi's criticism illustrates nicely the essential equivalence of efficiency-talk, cost-benefit-talk, and utility-maximization talk.

[13] Jane Jacobs, *Systems of Survival: A Dialogue on the Moral Foundations of Commerce and Politics* (New York: Random House, 1992).

> hours to mouth-blow and hand-cut a single glass. But then, our goal is not efficiency, but beauty.[14]

Of interest in both of these examples is the offhand assumption that efficiency itself can be a goal, or end. At first glance, this seems unproblematic. I may not think or visualize the concept of efficiency to myself when choosing the shortest route between the cleaners, the supermarket, and the bookstore. Nevertheless, I am certainly aiming to minimize the expenditure of time and energy as much as possible in doing my errands. If minimizing the expenditure etc. is the pedestrian concept of efficiency, and minimizing the expenditure etc. is what I am aiming at, then I am aiming at efficiency in the pedestrian sense. If I am aiming at it, then it is one of my ends.

Figure 3. Efficiency as a Goal of Action

[14] *The New York Times Magazine*, February 14, 1988, 3.

Although economists describe consumers as maximizing utility in their choices of commodity bundles, they do not ordinarily ascribe this to consumers as an end they intend to promote by making those choices. But this does not mean it is not one. For if, *ex hypothesi*, consumers thereby maximize utility in choosing the particular commodity bundles they choose, then maximizing utility is a conceptual or causal consequence of their choices. As such, it is either an intended or an unintended consequence. If it were an *un*intended consequence, then discovering that they had *failed* to maximize utility presumably would *not* lead them to revise their choices. But if consumers who chose commodity bundles that failed to maximize their utility would, upon discovering this, revise their choices, other things equal, then maximizing their utility is not an unintended consequence of their choices. Therefore it is an intended consequence of them, i.e. it is a goal or end. And therefore, the fact that economists do not recognize maximizing utility, i.e. efficiency, as an end does not mean it is not an end. Efficiency is an end, just as Jane Jacobs and the Waterford Crystal Company suppose.

However, the Waterford Crystal Company claims to *reject* this end for the sake of beauty. If the single end version of (U) is universal, it would then seem that the Waterford Crystal Company chooses beauty at the expense of rationality. But this is not obvious. Prereflectively, it is easy to imagine many alternatives to achieving an end efficiently – i.e. in a way that minimizes the expenditure of resources in its service – that are not prima facie irrational. I might opt for achieving an end expressively, elegantly, tastefully, honestly, diplomatically, traditionally, excitingly, gracefully, with panache, or in a way that preserves personal integrity. There is a prima facie *adverbial parity* between the concept of efficiency – i.e. of utility maximization – and those of beauty, elegance, honesty, etc. They each refer to conceptually distinct *styles or manners* of acting. Achieving my end efficiently is a manner in which I achieve my end – i.e. when I am feeling peppy and competent, just as is achieving an end diplomatically.

It is not hard to conceive a case in which these two manners of achieving an end might conflict. For example, the most efficient way of informing Clarence that his job application has been rejected may be to call him up and tell him just that. A more diplomatic, though less efficient way of conveying the same information would be to write Clarence a letter informing him who has been hired and thanking him for his interest. Other things equal, rejecting Clarence's job application in a way that does not hurt his feelings demands a more-than-minimal expenditure of resources in its service. That is, it demands that the goal of diplomacy override that of efficiency. If this overriding alternative end does not intuitively strike us as an obvious deviation from rationality, then we might need to invoke some more comprehensive

rationality principle in order to choose between the two. In Volume II, Chapter III I articulate one at length.

In addition to adverbial parity, there is also an *intentional parity* between the concepts of efficiency and the alternatives just mentioned. Achieving my end efficiently is not only a manner or style of achieving my end. Like other manners or styles, it is itself an end at which I may or may not aim – as is achieving my end tastefully, honestly, etc.[15] Let a *meta-end* be an end or goal regarding the style or manner in which I aim to achieve my ordinary ends, or *object-ends*. Meta-ends, on this account, are adverbial descriptions of action, of the sort just mentioned, that I aim to realize in acting to achieve my object-ends.

Orders of meta-ends may ramify as one considers in greater detail the style in which one wants to achieve some object-end. For example, Gladys may achieve her object-end of getting an education efficiently by going to college, and she may achieve this instrumental object-end in a way that tends to preserve her personal integrity by choosing among colleges with a socially progressive reputation. In this example, efficiency and the preservation of one's personal integrity are meta-ends.

Note also that the distinction between meta-ends and object-ends cuts across that between instrumental and final ends. A question as to the style in which one wants to achieve either instrumental or final ends may appropriately arise. And like object-ends, meta-ends, too, may be either instrumental or final in nature. Object-ends are, however, always instrumental to the achievement of meta-ends, whether these latter are instrumental or final in nature. Living the good life, for example, may be instrumental to the final meta-end of living in a graceful and aesthetically pleasing manner.

To say that I intend to achieve my end efficiently, or, alternatively, gracefully, does not imply that I must have such meta-ends consciously in mind when I act – anymore than I must my object-end. Let us say that I *minimally intend* to achieve some such end if, were I to discover that a particular action hindered this end, other things equal, this discovery would motivate me to refrain from performing it. Conversely, I *do not minimally intend* to achieve this end, if such a discovery would not motivate me to thus refrain from it. From now on, I shall use the word "intend" in this minimal sense. My main point is that achieving my end efficiently is itself a meta-end such that, if I were to discover that a particular action hindered it, it is an open question, dependent on context, whether this discovery would motivate me to refrain from performing it or not. So achieving my end efficiently is not a

[15] This distinction between adverbial and intentional parity corresponds to Jeremy Rifkind's distinction between efficiency as a method and efficiency as a value in his *Time Wars* (New York: Henry Holt and Co., 1987), Chapter 8.

meta-end with a special value-neutral and conceptually necessary status. It is one contingent value among others among which an agent may legitimately choose.

It may seem, however, that this conclusion ignores the conceptual distinction between adverbial intention descriptions and one particular group of adjectives and adverbs – call them *maximizing words* – that may always modify them, i.e. terms like "most," "more," "less," "greatest," "successfully," "optimally," and "maximally." We can always evaluate our actions in terms of *how fully, successfully, or maximally* they achieve the meta-ends of taste, honesty, efficiency, etc. This fact may suggest that rational action always involves maximizing something,[16] and that the seeming differences among these alternative meta-ends lie solely in the instrumental sources of utility they require us to maximize. If this is true, it means that efficiency or utility-maximization does have a special value-neutral and logically necessary status after all. For however we seek to realize our object- or meta-ends, it appears, we are acting rationally only if we are efficient in realizing them in precisely that way.

But it is a mistake to try to reserve this privileged position for the concept of efficiency or utility-maximization. If I achieve any end I intend to achieve efficiently by virtue of achieving it successfully or maximally, then I act efficiently merely by acting with deliberate intent. This makes the concept of utility-maximization vacuous. To see this, suppose Reginald is a mole in a local governmental bureaucracy, and that his assigned end is to impede the functioning of this bureaucracy as fully as possible. So he deliberately tries to achieve the explicit object-ends of this bureaucracy *in*efficiently. He achieves this meta-end by flooding himself and his staff with useless paperwork. Thus Reginald aims to achieve the meta-end of inefficiency itself efficiently, via the instrumental object-end of useless paperwork. But he may achieve this effectively obstructive flood of useless paperwork either efficiently, by convincing his superiors of its utility; or inefficiently, by printing up the forms himself and hoping his staff will use them. Assume Reginald chooses the former, instrumentally efficient strategy. Convincing his superiors of the utility of more paperwork is a further instrumental object-end which he may achieve efficiently, by arguing eloquently the advantages of extra paperwork;

[16] As D. M. Winch puts it, "the consumer is said to maximize utility, and utility is defined as that which the consumer attempts to maximize. This truism is completely general and cannot be false." (*Analytical Welfare Economics* (Harmondsworth: Middlesex, 1971), 17). Quoted in David Wiggins, "Weakness of Will, Commensurability, and the Objects of Deliberation and Desire," in Amelie O. Rorty, *Essays on Aristotle's Ethics* (Los Angeles: University of California, 1980), 260. David Gauthier also holds this view in "Economic Rationality and Moral Side-Constraints," *Midwest Studies in Philosophy III: Studies in Ethical Theory* (Minneapolis: University of Minnesota Press, 1978), 76-77.

or inefficiently, by making a few perfunctory and tactless remarks about their record-keeping practices. Again assume Reginald chooses the former, instrumentally efficient strategy. Arguing eloquently is a further instrumental object-end he can achieve efficiently, by marshalling and rehearsing his arguments in advance; or inefficiently, by relying on his native wit and ability to be fast on his feet.

Now suppose that here Reginald chooses the *latter* alternative. He realizes, of course, that he could argue his case for more paperwork before his superiors with far greater eloquence if he marshalled and rehearsed his arguments beforehand; and that he is likely to forget some of them, as well as deliver those he remembers with less polish, if he relies on his native wit. Nevertheless, rehearsing beforehand is just too boring. And since Reginald opts for excitement over efficiency when these two particular meta-ends conflict, he chooses to rely on his native wit. Now having made this choice, there may well be further choices of instrumental strategy to be made: whether to speak quickly or slowly, whether to use a Latinate or an Anglo-Saxon vocabulary, whether to wear a sports jacket or a three-piece suit, and so forth. In all such cases, Reginald may choose among these instrumental object-ends on grounds of efficiency, or on other grounds – of excitement, tastefulness, honesty, etc. At no point does rationality require him to opt for efficiency over these other meta-ends, if they are more important to him, *even if he thereby endangers his chances of efficiently sabotaging the bureaucracy's object-ends*. For although this is one of Reginald's long-term meta-ends, it may be perfectly rational for him to be unwilling to subordinate his personal style for its sake.

Now the efficiency expert may retort that Reginald thereby maximizes the instrumental meta-end of avoiding boredom and preserving his personal style by relying on his native wit; that he achieves this instrumental meta-end itself efficiently, by means of such reliance. But how does he maximize the instrumental object-end of relying on his native wit? By what means does he efficiently achieve *it*? The efficiency expert must reply that he maximizes the instrumental object-end of relying on his native wit simply by relying on it, or by performing any set of actions which relying on his native wit comprises. This is the "means" by which Reginald "efficiently achieves" this instrumental end. The efficiency expert's conception of achieving an end *E* by performing an instrumental action *A* thus conflates two cases: one in which *E* is a physically discrete causal consequence of *A*, in which case *E* can be achieved efficiently or inefficiently by means of *A*; and one in which *E* is merely an intentional redescription of *A*. In this second case, one "efficiently achieves" *E* by ascribing *E* to *A* as the object of its intention. All actions, in this sense, are "efficient means" to the ends they conceptually instantiate. So either efficiency

is one contingently valuable meta-end among others, or else it is vacuously equivalent to intentional action in general.[17]

Now suppose, *contra hypothesi*, that efficiency is a universal but *non*vacuous meta-end. Assume it is *universal* in that any basic action an agent performs in fact satisfies that agent's overriding desire to perform that action, and so maximizes his utility. If the meta-end of efficiency is universal, it has a special status that is not comparable to those of other meta-ends – honesty, taste, panache – that have adverbial and intentional parity. And since efficiency is a pervasive aim, it cannot conflict with any other meta-end. For however else one intends to act, and in whatever manner, one also intends to act efficiently.

Also assume the meta-end of efficiency is *nonvacuous* in that this is nevertheless *not* true by definition of "action." That is, performing a basic action that does *not* satisfy the agent's overriding desire to perform that action is a conceptual possibility. For example, suppose Edna has an overriding desire to speak from conviction. Yet when she says what she believes, the experience doesn't live up to her expectations. She feels that there must be some depth to her convictions she has not been able to plumb in speech. So she remains unsatisfied. Although she does speak from conviction, her utterance does not satisfy her overriding desire to speak from conviction. So Edna performs an action – uttering certain words – that does not satisfy her overriding desire to have uttered the very words she in fact uttered, and the

[17] Harsanyi explicitly embraces this latter alternative in *Rational Behavior and Bargaining Equilibrium in Games and Social Situations* (New York: Cambridge University Press, 1977), 17: "In effect, what we mean by 'rational behavior' is essentially behavior ... highly adapted, within the possibilities available to the person concerned, to successful achievement of his intended goals." Similarly, I. M. D. Little (*op. cit.* Note 8, 91, footnote 2) says, "The condition that an individual chooses a larger collection of goods is not, strictly, a postulate. It is an analytic proposition following from the meaning of an economic good. A larger collection of things is not necessarily a larger collection of goods." Also see Daniel Dennett, "Intentional Systems," *The Journal of Philosophy LXIII*, 4 (February 25, 1971), 87-106. The prevailing tendency to embrace the definitional equivalence of utility-maximization and intentional action confutes David Wiggins' claim that "[t]he statement that there is something the subject seeks strictly to maximize is not itself a definition, and must be allowed to take its chance with other empirical sentences....To defend it as a truism is to make into humbug everything that social scientists say in deference to Popper about falsification." (*ibid*., 260). I think Wiggins here misidentifies (U) as a testable empirical hypothesis, when in fact it functions in utility theory as a higher-level principle of interpretation (see Chapter IV, Section 4, below, for further discussion). In a footnote he dismisses this possibility, by declaring that the truistic representation of an agent's choices by indifference maps is not necessarily projectible. But of course if these indifference maps are interpreted in accordance with the truism in question, i.e. as a truism, then it is hard to imagine how they could fail to be.

very words that nevertheless ensure the efficiency of her action. Uttering that sequence of words fails to satisfy Edna's overriding desire to utter that very sequence of words. This example can be generalized to any overriding desire to perform a basic action that builds in an expectation of satisfaction relative to which the performance of the desired action falls short.[18]

But if performing a basic action that *does not satisfy* the agent's overriding desire to perform that very action is a conceptual possibility, then it is a conceptual possibility that one might not *minimally intend to satisfy* one's overriding desire to perform that very action. That is, were one to discover that an overridingly desired basic act *A* hindered efficiency – i.e. that performing *A* failed to satisfy one's overriding desire to perform *A*, one nevertheless would not refrain from performing *A*. So, to recur once more to the example, were Edna to discover that speaking from conviction failed to satisfy her overriding desire to speak from conviction, she nevertheless would not refrain from speaking from conviction. She might intend to speak from conviction anyway, perhaps from a sense of duty, or a need to preserve her integrity, or from habit, without regard to her desires or their satisfaction. So it is a conceptual possibility that one might not minimally intend to act efficiently.

However, this conclusion violates the above universality assumption, that however else one intends to act, and in whatever manner, one also intends to act efficiently. In this case, it seems, Edna both intends and does not intend to act efficiently. Similarly with any overridingly desired basic action that, in the performance, falls short of the satisfaction its performance was expected to give. The meta-end of efficiency – and therefore the single-end interpretation of (U) – cannot be both universal and nonvacuous without being inconsistent; indeed, self-contradictory.

This conclusion can be avoided by lifting the universality assumption from the single-end interpretation of (U). For if it is not true that one acts efficiently in performing whatever basic action one wants overridingly to perform, then there is no inconsistency in Edna's having on the one hand intended to act efficiently, and on the other failed to so act because she failed to satisfy her overriding desire to perform that act. There is no inconsistency

[18] In "Adaptive Utilities," (Allais and Hagen, *op. cit.* Note 2, 223-241), Richard M. Cyert and Morris H. DeGroot propose a method of deriving, from the gap between an agent's expected utility for some preference and her actual utility obtained as the result of satisfying it, her "adaptive or dynamic utility function" for the satisfaction of that desire. Cyert and DeGroot's proposal would remove the contradiction I describe only for those preferences in which multiple trials enabled the process of learning to occur. Of course, trial repetition does not *ensure* that learning occurs, even for a fully rational agent. Therefore Edna's disappointed expectation of desire-satisfaction does not depend on interpreting her action as a single-trial case. Being a "superstitious pigeon" is consistent with fully informed instrumental rationality under objective probability assignments.

here because not all rational ends need be efficiently achieved; and in particular, Edna's action of speaking from conviction may be rational even though it does not satisfy her desire to have done so. Edna's example shows that even in the case of basic actions, one may act inefficiently and still be rational in some wider sense that is not captured by the single-end interpretation of (U).

3. The Coherence Set Interpretation of (U)

3.1. Coherence

Next let us see whether this conclusion holds for a more comprehensive interpretation of (U). I describe as the *coherence set* interpretation of (U) the stipulation that utility is maximized when all of one's ends at a particular moment are ordered relative to one another. I conclude that here, too, universalizing the principle of utility-maximization implies that it is either vacuous or inconsistent.

In the coherence set interpretation, an agent maximizes utility by promoting as efficiently as possible the achievement of all of her ends conjointly. In this case the agent assigns a weight and a probability to each one of her ends, and trades off those of little weight or low probability against those with higher ones. She also discards those ends that are incompatible with others whose aggregate value is greater, as well as those which obstruct the achievement of others of greater aggregate value. She then schedules a plan for the achievement of those that remain. Many economists do not think of these calculations as actions a consumer performs. But they are nevertheless, and have costs and benefits just like all others. The costs and benefits of calculation themselves must be figured into the agent's calculations at the outset, so far as she is able to settle for herself the question as to whether these calculations are worth making; or whether it might rather maximize utility to organize all of her ends by repeatedly flipping a coin, perhaps, or according to a system of omens. Suppose she concludes that these calculations are worth making, and then proceeds to make them. Call the set of mutually coherent ends that results the agent's *coherence set*.

Now suppose Myrtle has the following two meta-ends, among others: (1) to achieve her object-ends simply; and (2) to achieve them efficiently in the single-end sense, i.e. with a minimum expenditure of resources in their service. Assume also that, as already argued in the preceding section, simplicity and efficiency have *prima facie* adverbial and intentional status. Then both of these meta-ends must enter into the calculus along with all the other ends Myrtle has. About the meta-end of efficiency, as with the meta-end of simplicity, Myrtle can and should calculate to what extent the aggregate value of each may outweigh the aggregate value of the other, and of object-ends that are instrumental to them. For example, Myrtle may need to settle

the extent to which minimizing expenditures may conflict with her wish to live a simple lifestyle. Simplicity may dictate weaving her own cloth, chopping firewood, drawing rain water from a bucket on the fire escape, growing her own fruits and vegetables, etc. How much time, energy and resources is Myrtle willing to devote to these activities for the sake of simplicity before they come into conflict with her desire to get things done efficiently? This is the kind of question that Myrtle's calculations should be able to answer.

Henceforth I shall use the term *efficiency* to refer to the single (meta-) end interpretation of (U), and the term *utility-maximization* to refer to the coherence set interpretation of (U) within which all of an agent's ends, both object-ends and meta-ends, including the meta-end of efficiency, must be situated. (U) in the coherence set sense remains a meta-end because organizing, balancing, and scheduling all of one's ends are themselves an object of goal-directed deliberative activity, i.e. they constitute an end. And (U) in this sense is a meta-end because it is a style or manner in which one may achieve each of one's object-ends, i.e. such that the achievement of some one object-end advances, or at least does not obstruct, the achievement of any of one's other object-ends.

The resulting conception of a coherence set appears to be *universal*. First, it applies to the totality of any agent's ends, regardless of content. Second, it subjects any particular end or meta-end, including that of efficiency, to the same cost-benefit analysis by which all of the agent's other ends must be mutually adjusted. It also thereby illuminates the sense in which utility-maximization in this universal, cost-benefit sense is *not* just one more contingently valuable meta-end, but rather does have a special value-neutral, logically necessary status. For when Myrtle asks herself at what point she should trade off the meta-end of efficiency against the meta-end of simplicity, she is really asking herself whether the cost of efficiency – i.e. sacrificing simplicity – may or may not outweigh its benefits; and whether, in fashioning her lifestyle, it really maximizes her utility to be efficient.

Since at bottom efficiency just is maximizing utility, this question may seem at first glance to have a paradoxical ring to it. But it is no more problematic than the question whether satisfying a certain desire itself is satisfying. In either case, there is no special difficulty about evaluating a lower-order value from the reflexive standpoint of that same value as a higher-order criterion. Efficiency as a meta-end that adverbially modifies an agent's achievement of various object-ends may be one such meta-end among many, all of which are subject to the higher-order regulating constraint of utility-maximization, itself an overriding meta-end of special status.

The resulting conception of a coherence set also seems to be *nonvacuous*, in that it is conceptually possible for an agent to violate it, i.e. fail to maximize utility, by failing to thus order all of her ends. She might include in the set an

end that conflicts with or obstructs others she deems more important. For example, Myrtle may find it very difficult to live simply, get things done efficiently, and hold down a full-time job. Or, to take another example, an agent might assign the greatest weight to an end with the lowest probability, such as winning the lottery, and subordinate all of her other endeavors to that one – thereby depriving herself of the resources necessary to achieve any of them, as the gambling addict does. So it would seem that the coherence set interpretation of (U) is not susceptible to the reproach of vacuous universality.

3.2. Nonvacuity

But now let us examine each of these features, i.e. universality and nonvacuity, at greater length. First assume universality and consider nonvacuity. The claim is that an agent's coherence set as just described is nonvacuous in the sense that it is possible for him to violate it, by failing to order all of his ends in the requisite way. But *is* it possible for him to thus fail to order his actual ends? Suppose that, having arrived at such a coherence set, he now proceeds to pursue an end not contained in the set, that conflicts with its ordering. Is he maximizing utility anyway, or is he not?

There are at least three possible answers to this question. A first is that he is not; that he is then acting irrationally, since he is, by hypothesis, not advancing all the ends contained in the set. This is the answer the utility theorist should give. Nevertheless it ignores the criterion according to which we were originally supposed to identify irrationality, namely failure to adjust all of one's ends so as to produce a coherence set. The case is one in which the agent is guilty of no such failure. He simply pursues an end not contained in the set he successfully ordered.

It is tempting to respond that if he pursues this end, then it is his end that then must be ordered relative to the set. But this does not follow. It is not difficult to imagine a case in which an agent pursues an end that is not his own. For example, wives traditionally have been expected to pursue their husbands' ends, regardless of what they thought about those ends, and have done so sometimes despite their own severe reservations or opposition to them. If an agent can pursue an end that is not his own, then an agent can pursue an end not contained in the coherence set of all his ends.

A second answer might be that if the agent pursues such an end, then *if it is* his end, it conflicts with the coherence set he has ordered. He has, therefore, violated that set, has thereby failed to maximize utility, and so has acted irrationally. But the fact that he is pursuing an end of his that conflicts with the set he ordered is evidence – indeed, for the revealed preference theorist, conclusive evidence[19] – that he has reorganized his priorities, reordered the set to incorporate the seemingly delinquent end, and indeed has ascribed to it

[19] I address revealed preference theory at length in Chapter IV, Sections 2 and 3, below.

overriding importance. So he *is* conforming to his coherence set, and his pursuit of this seemingly delinquent end promotes all of his ends after all. This answer saves the special, universal status of the utility-maximizing criterion. But it also implies that any end one pursues in action retrospectively satisfies the constraints of one's coherence set by reorganizing its priorities accordingly. This makes the coherence set vacuous, for any action one takes can be made to satisfy its requirements.

A third answer might be to concede that the delinquent end the agent pursues is an end, deny that it is *his* end, and deny also that pursuit of it counts as a genuine action. Here the thought would be that although he behaves intentionally in virtue of aiming at an end, the end he aims at is not an end he desires to obtain; so he merely "goes through the motions" of acting, without being motivated by the conative resolve that desire ignites. But this third answer would also underwrite the conclusion that his coherence set is vacuous. For it implies that any action that does not conform to it is not really an action at all. This means, in turn, that an agent cannot rationally regard any end he pursues as thwarting its constraints, consistently with regarding it as his end. He must discount any action he performs that appears to thwart its constraints either as third-personal, behavioral evidence that he has altered his coherence set to accommodate it; or else as mere physical behavior that is not, in fact, a genuine action. But this just seems mistaken.

Suppose, for example, that Sylvester must order the object-ends of being a dentist, being a poet, and making lots of money, and invokes the meta-end of utility-maximization to do so. Having decided that being a poet is incompatible with making lots of money whereas as being a dentist promotes it; and, moreover, that being a rich dentist would make him happier overall than being a poet, Sylvester then finds himself writing poetry, attending readings, publishing a little magazine, and neglecting his dental practice. He is deeply troubled about this behavior, and regrets in advance the many cavities he will not fill, dollars he will not make, and cruises he therefore will have to forego.

Obviously such cases of internal conflict require their own complex analysis. But there are two claims we probably ought not to include in such an analysis. First of all, it is not open to us to conclude that Sylvester's pursuit of poetry is not a genuine action at all. Writing poetry, attending readings, and publishing a little magazine are definitely actions, *if any behavior is*. Second, if he clearly recognizes that he is sacrificing happiness for the sake of his poetry (perhaps he even hopes that his suffering will improve it), it is not open to us to conclude that Sylvester's coherence set has changed to accommodate his poetry-seeking behavior – at least not without inviting the threat of vacuity. For recall that Sylvester organized that set according to the meta-end of utility-maximization. By sacrificing his happiness in order to write poetry,

Sylvester clearly fails to achieve that meta-end, *if any action can*. If no action can, then the vacuity of that meta-end follows immediately.

The notion of a coherence set of ends organized to maximize utility overall is defective as a universal criterion of rationality, because it assumes that we can always make an ordinal trade-off between those ends the achievement of which maximize utility and others whose utility costs are too great. But this assumption is false. Ends that are discarded from the coherence set on grounds of inferior utility do not necessarily disappear, if their qualitative character is sufficiently compelling. They may remain sources of intense longing and regret – intense enough to motivate action in their service, without upsetting the original ordering of the set. Indeed, their qualitative importance may increase, even as their ordinal rankings decrease. Therefore, an agent may rationally choose to forego utility-maximization for the sake of those ordinally inferior but qualitatively compelling ends, if they are compelling enough. To reply that an end that is sufficiently qualitatively compelling to motivate action must be ordinally overriding as well merely repeats the error of reasoning I am targeting, for it begs the question of whether or not qualitative superiority is reducible to ordinal superiority. Sylvester's choices suggest that it is not.

Thus an agent may choose rationally to live in a way that fails to maximize utility, and to accept his consequent unhappiness or dissatisfaction, if other considerations – for example, giving expression to his deepest impulses – are more qualitatively important (*not*: "more satisfying") to him. To then retort that if these other considerations really are more important to him then he has maximized utility after all is revert to the vacuous single-end interpretation of (U), in which any action one performs maximizes utility by definition. The vacuity of the coherence set can be avoided only by denying its universality.[20]

[20] H. A. Simon's modifications of the utility-maximization model of rationality (*op. cit.* Note 8) seem to me unsuccessful in circumventing the worries I have raised because, unlike Liebenstein's theory of "selective rationality" (*op. cit.* Note 8) which attempts to reformulate (U) in terms of a basic concept of "trying" or "effort", Simon's notion of "satisficing" is equally susceptible to the charge of vacuity. Indeed Simon comes close to acknowledging as much when, in discussing changes in an agent's "aspiration level" as definitive of a satisfactory alternative, he states that "[s]uch changes in aspiration level would tend to bring about a 'near-uniqueness' of the satisfactory solution and *would also tend to guarantee the existence of satisfactory solutions. For the failure to discover a solution would depress the aspirational level and bring satisfactory solutions into existence.*" (italics added; "A Behavioral Model of Rational Choice," 111) Simon's satisficing agent would seem to be incapable of frustration, disappointment, or fear of failure. In fact her dissatisfaction level would be so low, and her contentment level so easily reached, that her motivation for acting in any way at all is obscure.

3.3. Universality

Now again assume, as in the single end interpretation of (U), that an agent's coherence set is *non*vacuous, and consider further its universality. This means that on the one hand, it is conceptually possible for an agent to violate the set by pursuing an end not contained within it. On the other, its ordering subjects all meta-ends, including that of efficiency, to the same cost-benefit analysis by which all of an agent's object-ends must be mutually adjusted, regardless of particular content. So although this is not true merely by definition of having an end or of maximizing utility, utility-maximization in this sense is *not* just one more contingently valuable meta-end, but rather does have a special necessary and universal status.

If utility-maximization is a necessary and universal meta-end, then it is an *absolute* meta-end. First, it is *permanently superior in ranking* to all of an agent's other meta- and object-ends. Like moral side-constraints on action,[21] utility-maximization is, first of all, an intentional object (conscious or otherwise) of goal-oriented behavior, i.e. it is an end. Second, it is an end that is *not subject to revision or sacrifice* for the sake of otherwise realizing any further object- or meta-ends. Rather, one must sacrifice, revise, or reschedule these other ends in order to maintain conformity to the constraints one's coherence set imposes, just as (U) requires. Failure to make such revisions results in violation of the set.

Moreover, utility-maximization in this universal sense is a *final* meta-end, in that any considerations invoked to justify its imposition must be noninstrumental in nature. So I cannot convince you to organize all of your ends into a coherence set that maximizes utility by pointing out that it maximizes utility to do so. For this would make (U) on the coherence set interpretation either instrumental to some higher-order (U) on some other interpretation; or else redundant. Similarly, I cannot justify your satisfying a certain desire by pointing out that it satisfies that desire to do so.

In Chapter II.2 we have already previewed Chapter VIII.2's argument below, that one may ramify orders of desire infinitely in recursive acts of self-evaluation, without meeting the requirement of rational justification. But quite independently of that argument, we can already see that invoking any such higher-order criterion of evaluation to rationally justify itself is merely redundant. Hence Chapter II's conclusion regarding desire applies here to utility-maximization: whatever the considerations are that justify organizing one's ends in order to maximize utility, these must be independent of utility-maximizing considerations themselves. But within this model of rationality, utility-maximizing considerations are the only considerations available. So according to the coherence set interpretation of (U), any further

[21] of the sort discussed by Robert Nozick in Chapter III of his *Anarchy, State and Utopia* and David Gauthier (*op. cit.* Note 1).

considerations, whatever they are, must be arbitrary from the point of view of rational justification.[22]

So, to recur to the example, suppose Sylvester wants to know whether it is best for him to organize his ends according to (U) in the first place. That principle itself can afford him no answer. He may seek one outside the constraints of the utility-maximizing model of rationality – such as that

(E) If a rational agent maximizes utility, he expresses himself.

But if (E) is, indeed, external to the model, then either it justifies (U) by invoking a more basic value – self-expression – to which (U) is in fact subordinate; or else it has no authoritative status relative to (U), and so cannot be invoked as a reason for adhering to it. If (E) is internal to the model, on the other hand, then the costs and benefits of following it must be instrumentally calculated just like any other. Finally, if (U) is so understood to presuppose or imply (E), then (E) is not independent of utility-maximizing considerations after all. Thus invoking a principle such as (E) as the "further consideration" that would purport to justify (U) itself would be unsuccessful. Either (U) is a first principle, or it is not.

If utility-maximization is an absolute and final meta-end, it cannot be abandoned if its opportunity costs relative to other final ends seem too high. For by hypothesis, utility-maximization is not the kind of end to which such costs and benefits themselves can be assigned. Suppose Sylvester carefully and reflectively decides that he would be happier overall being a rich dentist than being a poet. Then according to this model, he *cannot* then go on to reject the meta-end of being happier overall on the grounds that it obstructs expression of his deepest creative impulses to attain it. That would be to reject the very criterion relative to which expressing his deepest impulses was evaluated and rejected. If Sylvester decides that expressing his creative impulses is more important, then on this view that is what makes him happiest overall. That is what maximizes his utility.

So utility-maximization is not only an absolute, final meta-end, the status of which is arbitrary with respect to rational justification. In addition, it is not

[22] Indeed this implication is explicitly embraced by the progenitor of principle (U) when he argues that

> 'Tis not contrary to reason to prefer the destruction of the whole world to the scratching of my finger. 'Tis not contrary to reason for me to chuse my total ruin, to prevent the least uneasiness of an *Indian* or person wholly unknown to me. 'Tis as little contrary to reason to prefer even my own acknowledg'd lesser good to my greater, and have a more ardent affection for the former than the latter.

(David Hume, *A Treatise of Human Nature*, Ed. L. A. Selby-Bigge (Oxford: Clarendon Press, 1968), 416). In Chapter XIV, I argue that Hume is not merely being provocative in this and comparable passages, i.e. that Hume really is a Humean.

an end about which further rational deliberation is possible. Of course this does not mean it cannot function as a methodological criterion for identifying and evaluating rational action from the third-person perspective. But it does mean that it cannot be an ascribable meta-end that any utility-maximizing agent might rationally and deliberatively choose to adopt. It is either an absolute, final, and rationally arbitrary meta-end within the rationality model of utility-maximization, or it is not an end at all. It would be unfortunate indeed if *this* were the reasoning that motivated some economists to deny that (U) is an end.

But an absolute and final meta-end that is by definition incapable of entering into a utility-maximizing agent's cost-benefit analysis cannot be ranked relative to her other final ends, *not even as superior to all of them*. For to ascribe to it this superior ranking implies that, all things considered, it has *lower* opportunity costs than any alternatives. And this presupposes the contingent dispensability of (U) we have just seen is excluded by stipulation of its universality. Hence utility-maximization is not just a rationally arbitrary, absolute, final meta-end; it is a conceptually inconsistent one. For it both is and is not superior in ranking to all other ends, so both is and is not absolute, and so both is and is not universal. Therefore it is not a meta-end that any utility-maximizing agent could consistently intend, even minimally, to adhere. It seems, then – here as in the single-end interpretation, that assuming (U) to be both nonvacuous and universal implies that it is inconsistent. (U) can be made consistent only if it is either vacuous or limited in its scope of application.

I take it that these conclusions give us some reason to rethink the claim that (U) is universal. For of course people do sometimes carry out their intentions to accomplish things efficiently, and to maximize utility in all of their projects. The argument has not been that utility-maximization is an inherently inconsistent end. Rather, it is inherently inconsistent when conceived as an *absolute final* end. Utility-maximization as an overriding value could not be both universal and nonvacuous in its application, for in that case it would be conceptually inconsistent. As soon as we acknowledge that utility-maximizing considerations might be compared with other contingently valued meta-ends according to completely different and extrinsic rationality criteria and ranked or rejected accordingly, the inconsistency disappears. If this strikes you as a reason to reject its claim to universality, then you must view the value of rational consistency as overriding it. This is the extrinsic rationality criterion I shall try to defend at length in Volume II.

4. Three Interpretations of "Utility"

So far I have argued that, in order to avoid the Scylla of vacuity and the Charybdis of inconsistency, (U) must be understood as contingent and restricted in its scope of instantiation. In discussing the single end and

coherence set interpretations of (U) in Sections 2 and 3 above, I have deployed an uninterpreted concept of utility itself that is loosely interchangeable with commonsense concepts of happiness, desire-satisfaction, etc. It may seem that my arguments have depended on this uninterpreted concept of utility, such that they would collapse if this concept were specified to denote one particular set of intentional conditions rather than some other. But I now go on to reject the suggestion that the implied vacuity or inconsistency of (U) depends on such an uninterpreted concept of utility, and fails when some particular interpretation is supplied.[23] In this section I argue that it does not matter whether the concept of utility is interpreted phenomenologically, psychoanalytically, or behaviorally. The problem remains: If (U) describes an end that agents always have, i.e. a universal end, then it is vacuous. If the utility theorist tries to maintain universality while denying vacuity, then we may simply repeat the reasoning of Sections 2 and 3 and conclude to inconsistency.

4.1. The Phenomenological Interpretation

The classical concept of utility was understood to refer to occurrent happiness, pleasure, or the satisfaction of desire[24] as a conscious mental state, or disposition to have such states.[25] Call this the *phenomenological interpretation* of the concept of utility. As we have already seen in Chapter II.2.1, happiness, pleasure and desire-satisfaction are notoriously nonequivalent. But for purposes of the present argument we lose nothing by regarding them as more or less interchangeable.[26] On the phenomenological interpretation, a fully

[23] So far as I know, this suggestion first appears in Ward Edwards, *op. cit.* Note 2, 382.

[24] The traditional definition of utility as happiness or pleasure is to be found in Sidgwick, *The Methods of Ethics* (New York: Dover, 1966), Book I, Chapter IV; Book II, Chapters I-III; Book III, Chapter XIV; Book VI, Chapter I. Also see Jeremy Bentham, *Introduction to the Principles of Morals and Legislation*, Ed. J. H. Burns and H. L. A. Hart (London: Athlone, l970), Chapter I, Sections 1.-2. Richard Brandt discusses the merits of the "happiness' versus the "desire" theory in his *A Theory of the Good and the Right* (Oxford: Clarendon Press, 1979), Chapter XIII. Also see his "The Concept of Welfare" (unpublished paper, 1980).

[25] Brandt's own definition is dispositional. See his *A Theory of the Good and the Right*, *ibid*. Chapters II.1 and XIII.2; and Richard Brandt and Jaegwon Kim, "Wants as Explanations of Actions," in N. C. Care and C. Landesman, Eds. *Readings in the Theory of Action* (Bloomington, Ind.: Indiana University Press, 1969), 199-213.

[26] But see Brandt's analysis of the different implications for utilitarianism of each in *A Theory of the Good and the Right, ibid*. Wayne Davis analyses occurrent happiness in terms of desire-satisfaction in "A Theory of Happiness," *American Philosophical Quarterly 18*, 2 (April 1981), 111-119; and identifies occurrent happiness with pleasure in "Pleasure and Happiness," *Philosophical Studies 39* (1981), 305-317.

rational agent performs actions that multiply and intensify these states as fully as possible. Thus particular final as well as instrumental ends are understood as instrumental to the further, ultimate end of maximizing utility.

The stipulation of happiness or desire-satisfaction as the ultimate end does not in all cases imply that all of an agent's particular ends must be instrumental to it. Instead, some might be constitutive of it, and so could be identifiable final ends in their own right.[27] But in utility theory, there is an implicit distinction between structure and intention that requires this inference. Let *a, b, c, ...* be *structurally constitutive* of *X* if *a, b, c, ...* are in fact always found together, and together constitute *X*. Let *a, b,* or *c, ...* be *structurally instrumental* to *X* if *X* is a causally or conceptually distinct consequence of *a, b,* or *c*; and not vice versa. Let *a, b,* and *c* be *intentionally constitutive* of *X* if to intend, desire, or believe *a, b, c, ...* is to intend, desire or believe *X*. And finally, let *a, b,* or *c* be *intentionally instrumental* to *X* if one intends, desires or believes *a, b,* or *c* only if one believes that *a, b,* or *c* results in *X*.

Now it may be that the achievement of certain ends is inextricably linked with a state of happiness or desire-satisfaction, in that their achievement is always in fact accompanied by it. Experiencing deep and satisfying friendships, or fulfilling work, or a work of art, may have this character, whereas driving a hard bargain, having a well-paying job, or listening to an edifying lecture may not. Ends that are inextricably linked with the experience of happiness or desire-satisfaction are, to be sure, structurally constitutive of those mental states.[28] Nevertheless if their only intentional function is to engender these states; if one consciously works to achieve these ends only if and because they result in these states, and not for any other reasons (such as that they are intrinsically valuable), then they are intentionally instrumental to them. And this is the role that classical utility theory assigns to all such ends. They are all instrumental to the maximization of utility, understood as happiness, pleasure, or desire-satisfaction. Hence classical utility theory

[27] See W. F. R. Hardie's distinction between inclusive and dominant ends in "The Final Good in Aristotle's Ethics," *Philosophy XL* (l965), 277-295.

[28] I take Sidgwick to be *denying* this point when he says that "if I in thought distinguish any feeling from all its conditions and concomitants – and also from all its effects on the subsequent feelings of the same individual or of others – and contemplate it merely as the transient feeling of a single subject; it seems to me impossible to find in it any other preferable quality than that which we call its pleasantness, the degree of which is only cognizable directly by the sentient individual." (*The Methods of Ethics, op. cit.* Note 24, Book II, Chapter II, Section 2, p. 128).

assigns to (U) a universal and logically necessary status. And we have already seen that this means it is either vacuous or inconsistent.[29]

4.2. The Psychoanalytic Interpretation

Utility theorists sometimes try to meet this charge by appending to (U) a theory of unconscious desires. They reason that actual agents do not invariably maximize utility merely by acting intentionally, because actions are sometimes motivated by unconscious, destructive desires that may cause one a great deal of conscious unhappiness or dissatisfaction. Hence though (U) is true for fully rational agents, it is not true for conflicted, ambivalent, self-destructive, or self-deceived actual agents. Hence it cannot be vacuously true. Call this the *psychoanalytic interpretation* of the concept of utility.

Let us grant, for the sake of argument, the distinction between fully rational and imperfectly rational agents that I claimed in the introduction to this chapter to be irrelevant to my argument in this section. Even if we do so, this interpretation does not have the implications its proponents claim. Stipulating the existence of unconscious desires whose satisfaction thwart conscious ends implies unconscious ends they do not thwart but rather achieve. Then conscious actions and the ends they promote become instrumental means to the achievement of those unconscious ends.[30] And the

---

[29] In this respect, classical utility theory seems mistaken on purely common-sense psychological grounds. Happiness or pleasure may be merely a contingent consequence or side-effect of an end we deliberately adopt, rather than that to which all our ends are intentionally instrumental, as Bishop Butler argues (*Fifteen Sermons*, Sermon XI, 415; reprinted in *The British Moralists 1650-1800, Volume I: Hobbes-Gay*, Ed. D. D. Raphael (Oxford, The Clarendon Press, 1969). For example, gratification may be a valuable side-effect of personal integrity, but individuals may strive to achieve this end even when no such gratification is anticipated. Reliance upon common-sense distinctions among our mental states, and consequent application of the terms "utility," "happiness," or "desire-satisfaction" to some of them and not others, enables us to both retain the conceptual resources for distinguishing expected utility-maximization from other ends of action, and thus pick out the full range of mental phenomena a social theory is concerned to explain. (David Lewis makes much the same point in "Radical Interpretation," *Philosophical Papers, Volume I* (New York: Oxford University Press, 1983), 110, when he remarks, "If our interest is in the philosophy of mind and of language, then the pursuit of ontological parsimony seems to me an unnecessary distraction" – without, I think, seeing the implications for his own use of the belief-desire model of action in that discussion.) In this case, (U) holds only under certain contingent circumstances that may or may not obtain for an agent. But she may act with full rationality nevertheless.

[30] But see Peter Alexander, "Rational Behavior and Psychoanalytic Explanation," in Care and Landesman, for a different view. Theodore Mischel defends Freudian explanation against Alexander's criticism (misguidedly, I think) in "Concerning Rational Behavior and Psychoanalytic Explanation," *Mind 74* (1965), 71-78. A more moderate defense is provided by Robert Audi, "Psychoanalytic Explanation and the Concept of Rational

conscious utility such actions fail to maximize become mere opportunity costs that are by definition outweighed by the unconscious utility they succeed in maximizing.

Moreover, the point has been made often that there are no firm theoretical constraints on when we are justified in invoking unconscious desires to explain action, nor even on what those desires must be.[31] This means that whenever an action appears to be destructive or self-defeating for the agent who performs it, an unconscious desire it satisfies can always be found. So interpreting (U) to include unconscious as well as conscious desires does not circumvent the charge of vacuity. Quite the contrary.

4.3. The Behavioral Interpretation

The phenomenological and psychoanalytic interpretations of the concept of utility both rely on the background concept of a phenomenal mental state. It may seem that this background concept is to blame for failing to block the charge of vacuity. Of course a mental state-conception of utility is vacuous, the argument might go. Since we never experience another's conscious (or our own unconscious) motivation first hand, we are free to speculatively attribute to an agent any conscious or unconscious motive we like in order to explain his behavior.[32]

Action," *The Monist 56* (1972), 444-464. Alexander's thesis is augmented by Harvey Mullane, "Psychoanalytic Explanation and Rationality," *The Journal of Philosophy LXVIII*, 14 (1971), 413-426.

[31] To my knowledge, the first argument to this effect is to be found in B. F. Farrell, "The Criteria for a Psychoanalytic Interpretation," *Proceedings of the Aristotelian Society, Supplementary Volume XXXVI* (1962). Also see Karl Popper, *Conjectures and Refutations: The Growth of Scientific Knowledge* (New York: Harper and Row, 1963), 37-38; Frank Cioffi, "Freud and the Idea of a Pseudo-Science," in Robert Borger and Frank Cioffi, *Explanation in the Behavioral Sciences* (Cambridge: Cambridge University Press, 1970), 471-499; Adolph Grünbaum, "How Scientific is Psychoanalysis?" in Raphael Stern, Louise S. Horowitz, and Jack Lynes, Eds., *Science and Psychotherapy* (New York: Haven, 1977); "Is Freudian Psychoanalytic Theory Pseudo-Scientific by Karl Popper's Criterion of Demarcation?" *Americal Philosophical Quarterly XVI,* 2 (April 1979), 131-141; "Epistemological Liabilities of the Clinical Appraisal of Psychoanalytic Theory," *Nous XIV*, 3 (September 1980), 307-385; Richard Nisbett and Timothy Wilson, "Telling More than We Can Know: Verbal Reports on Mental Processes," *Psychological Review LXXXIV* (1977), 231-259; Edward Erwin, "The Truth about Psychoanalysis," *The Journal of Philosophy LXXXVIII*, 10 (October 1981), 549-560.

[32] This argument appears explicitly in Joel Feinberg, "Psychological Egoism," in Joel Feinberg and Russ Shafer-Landau, Eds., *Reason and Responsibility: Readings in Some Basic Problems of Philosophy* (Belmont, Cal.: Wadsworth Publishing Company, 1998), 493-505; and in James Rachels, *The Elements of Moral Philosophy* (New York: Random House 1986).

This argument explains the vacuity of utility-ascriptions by the inaccessibility of mental states. It reasons that this inaccessibility furnishes the license to arbitrarily conjecture mental states of third-personally observed agents or actions, and so to ascribe utility-maximizing motives in a similarly arbitrary manner. The implication is that a physical state-conception of utility would avoid these difficulties, because a physical state is an interpersonally accessible state that restricts speculative motivational explanations to what is third-personally observable. Thus it motivates a *behavioral interpretation* of the concept of utility, the theory of revealed preference. According to revealed preference theory,[33] utility rankings – preferences – are revealed in observable behavior. Any action an agent performs expresses his preference for that result actually achieved, and/or some finite set of its consequences. If one chose or preferred that end one actually achieved, then all of one's behavior is by definition fully intentional, and one always chooses to act as one most prefers to act.[34]

But if every action expresses the overriding preference to have performed precisely that action, then citing that preference can provide no independent explanation of why that action, rather than some other, was performed; this is Nagel's concept of a motivated desire, discussed at greater length in Chapter VII.2.3 below. Instead we must seek an account of why the agent had that preference. Such an account might appeal to environmental and biological influences; the agent's experiences, values, and beliefs; social and historical determinants; impulses, cravings, and occurrent drives (i.e. Nagel's unmotivated desires); and so forth. Such an account will illuminate some of the causal connections between these factors and the agent's actual behavior. But it also eliminates the concept of preference revealed in behavior – i.e. of motivated desires – as an intervening explanatory variable. At least this oversimplified formulation of the theory would seem to hasten the conclusion to vacuity rather than sidestep it.

---

[33] The basic idea of revealed preference theory is in Frank P. Ramsey, "Truth and Probability," in *The Foundations of Mathematics and Other Logical Essays,* Ed. R. B. Braithwaite (London: Routledge and Kegan Paul, 1950), 157-198. P. A. Samuelson first formulates it explicitly in "A Note on the Pure Theory of Consumer Behavior," and "A Note on the Pure Theory of Consumer Behavior: An Addendum," *Economica 5* (1938), 61-71 and 353-4. Also see I. M. D. Little, *op. cit.* Note 8, 90-99. Little later anticipates the vacuity problem discussed here, but apparently without seeing its implications for the theory of revealed preference. See his *Critique of Welfare Economics* (New York: Oxford University Press, 1970), Chapter II, esp. 21-22. Amartya Sen states and critiques these implications in "Behavior and the Concept of Preference," *Economica 40* (1973), 241-259; and "Rational Fools: A Critique of the Behavioral Foundations of Economic Theory," *Philosophy and Public Affairs 6*, 4 (1977), 317-44.

[34] See Tibor Scitovsky, *The Joyless Economy*, p. xi, *op. cit.* Note 1.

Moreover, Sen has argued that the theory of revealed preference has not succeeded in detaching the concept of utility from inaccessible mental states. For the notion that an individual expresses or reveals his expected utility rankings in his behavior presupposes that there is, indeed, some internal mental state his behavior reveals.[35] Then it is an open and empirical question whether his observable behavior actually does express that state, or whether it is motivated by covert, strategic ends. If it is covertly motivated, there exists an instrumental relationship between his overt behavior and the covert ends it promotes. If one of his preferences is to conceal his true preferences, then he maximizes utility by expressing false preferences in his overt behavior. Therefore, either the agent's behavior itself satisfies his preferences – thus maximizing utility constitutively, or else it is instrumental to the satisfaction of covert preferences – thus maximizing utility instrumentally. Structurally, the behavioral interpretation of utility has the same infelicitous consequences as the psychoanalytic interpretation. The following chapter pursues further some of the infelicities attendant on the behavioral interpretation of utility-maximization.

[35] Amartya K. Sen, "Behavior and the Concept of Preference," *op. cit.* Note 32.

Chapter IV. The Utility-Maximizing Model of Rationality: Formal Interpretations

In this chapter I consider and evaluate some more formal versions of the claim to universality of the utility-maximizing model of rationality (U), as formulated in Section 1 of Chapter III. Readers who are uninterested in technical exposition are invited to skip directly to Sections 4 and 5 below, which offer conclusions pertinent to the discussions of both chapters. Section 1 criticizes the attempts of Von Neumann-Morgenstern, Allais, and others to extend the behavioral and phenomenological interpretations respectively of (U) into interpersonal comparisons of utility. This extension is usually understood to be required merely for deriving a social utility function from individual ones. But the viability of third-personally observable ascriptions of utility, and hence of the behavioral interpretation of (U), themselves presuppose an interpersonally shared criterion of utility, and therefore the ability to make interpersonal comparisons. Even Allais' phenomenological interpretation of (U) depends on an unexamined conception of veridically observable speech behavior. Hence my argument, that interpersonal comparisons – and therefore the interpersonally communicable utility-ascriptions based on them – are impossible whether utility states are empirically observable or not, has unfortunate implications for both interpretations.

Section 2 contends that the Ramsey-Savage consistency constraints on preference rankings by themselves do not succeed in both rescuing the utility-maximizing model from vacuity and preserving its universality, because they do not successfully exclude preferences that are, from an external perspective, identifiably logically inconsistent; and that excluding them by fiat does not solve the problem. I suggest – but defer to Volume II extended defense of the thesis – that constraints of logical consistency would rescue (U) from vacuity. I similarly defer to that volume extended discussion of McClennen's concept of resolute choice as in effect supporting my analysis. I issue a promissory note, to be redeemed in Volume II, Chapter III, as to how such constraints of logical consistency on preferences might be symbolized, using the traditional Boolean connectives and some familiar conventions of predicate logic, so as to formally exclude cyclical rankings. Finally, I defer to that discussion an extended argument for the thesis that (U) is a special case of a different conception of logically consistent choice that is much broader in scope. Hence my discussion in Sections 2 and 3 merely lays the critical groundwork for several substantive suggestions I make later as to how (U) might be rethought so as to avoid some of the problems I raise here.

Unlike most formal decision theory, which basically ignores the traditional Boolean connectives, Jeffrey-Bolker expected utility theory[1] does not. It uses them to construct preference alternatives comprising strings of weighted and probabilistically defined propositions and complex gambles among them. But it leaves undisturbed the conventional connectives imported from mathematics (">," "≥," "=,") for ordering those complex preferences themselves. This leaves moot the question whether or not strictly logical interrelationships among them also obtain. So it may seem that in general, the canonical notation and axiomatic formulation of decision theory place it outside the purview and constraints of classical logic. However, the mere fact that formal decision theory in its canonical symbolization does not recognize the constraints of classical logic would not seem sufficient grounds for inferring that those constraints do not apply. Similarly, the fact that we cannot seem to symbolize logically the inconsistency involved in a cyclical ranking does not suffice to infer that no logical inconsistency is present. The question whether or not a cyclical ranking violates the law of noncontradiction is not in theory unanswerable. In Section 3, I merely raise this issue for discussion, by showing some of the commonsense ways in which a cyclical ranking certainly does seem logically inconsistent, even though the canonical notation of formal decision theory does not allow us to express this. I defer to Volume II, Chapter III a full and detailed treatment of this topic, including some suggestions as to how this notation might be modified so as to reveal its subordination, not only philosophically but also formally, to the requirements of logical consistency.

Sections 4 and 5 argue that thus relativizing the Humean model renders it, like the maximin principle, contingent with respect to the requirements of rational action more generally understood. Only after reaching this conclusion do I address the metaethical status of the utility-maximizing model of rationality. Most of this chapter looks at the explanatory reach of this theory *without regard to its metaethical status as normative or descriptive within any particular discussion*. The question whether a theory is explanatory or not can be answered independently of the question whether it has a normative or a descriptive metaethical status. The metaethical status of any principle is fully exhausted by specifying the relation between two descriptive versions of it: that which describes actual behavior and that which describes ideal behavior.

---

[1] See Richard C. Jeffrey, *The Logic of Decision*, Second Edition (Chicago: University of Chicago Press, 1983), especially Chapter 9; Ethan D. Bolker, "A Simultaneous Axiomatization of Utility and Subjective Probability," *Philosophy of Science 34* (1967), 333-340; and Bolker, "An Existence Theorem for the Logic of Decision," *Philosophy of Science 67* (2000), S14-S17. I discuss the role of the indifference relation in the Jeffrey-Bolker representation theorem in Volume II, Chapter III.7.

A theory can be both explanatory and normative if it explains the behavior of an ideal agent who sets a standard we are exhorted to emulate.

Section 4 considers the normative implications and conceptual inconsistencies of (U) as a limiting ideal, in order to demonstrate why it cannot be defended as a higher-level, prescriptive principle of interpretation that is therefore immune to Popperian falsification requirements. Yet even if the utility-maximizing model is a *bona fide* explanatory theory of necessarily limited scope, utility theory *can be* no more or less prescriptive and value-laden than moral theory. Section 5 therefore concludes with a brief comparison between Kantian moral theory and utility theory as two theories of value that compete for foundational status, each subordinating the other to itself, and between which we must choose to explain the data of human behavior. The utility-maximizing model is argued to be at best a contingent normative theory that loses out to a Kantian model of rationality for primacy within our conceptual scheme. I develop the thesis that a Kantian moral theory is a descriptive and explanatory theory to which we bear a special relation in Volume II, Chapters V.5 and IX.

1. Interpersonal Comparisons of Utility

The behavioral interpretation of (U) was intended to be a theory of consumer behavior freed from dependence on the concept of utility interpreted as an inaccessible mental state. The underlying reasoning was that utility states would be accessible if and only if they were empirically observable from a third-personal perspective. We could then meaningfully compare them, and ascribe utility rankings to an agent's ends on the basis of those empirically based comparisons. Since the actual values of each of those ends would be interpersonally accessible, the resulting utility-ascriptions would be empirically confirmable and so nonvacuous. But utility-ascriptions to third-personally observable behavior can be empirically confirmable only if there is some interpersonally shared criterion of utility relative which such utility-ascriptions can be confirmed. That is, the feasibility of third-personally observable ascriptions of utility – and so of the behavioral interpretation of (U), the theory of revealed preference – presupposes the ability to make interpersonal comparisons of utility. Unfortunately we have no such ability.

The problem of interpersonal comparisons ordinarily arises in the context of how to extract a social utility function from individual ones. But it is not unique to that context. My ascription of utility-maximization to your behavior requires a similarly rule-governed relationship between my criteria for assessing utility-maximization and your purported manifestation of it. That is, it requires that the type and degree of utility-maximization I see in your behavior be the same as the type and degree of utility-maximization you see your behavior as actualizing; that we share in common a quantitative standard of utility-maximization by which both my judgment and your

behavior can be calibrated. I now argue that interpersonal comparisons, and so the utility-ascriptions based on them, would be impossible whether utility states were empirically observable or not.[2] Hence their accessibility would ensure neither their empirical confirmability, nor, therefore, the nonvacuity of those utility-ascriptions.

1.1. The Social Utility Function

Interpersonal comparisons of utility are what we need to be able to make in order to derive from the measurement of individual utilities a social utility function representing the combined utility functions of individuals.[3] The simplest such function would be additive: all individual utility functions would be summed to total the social utility function. Other possibilities would include finding the average or median, or a representative utility function of all individuals.

Measuring the happiness or utility level of a single individual would seem to present no obvious difficulty. We simply ask her to make consistent pairwise comparisons among given available options F, G, and H such that F > G (= F is strictly preferred to G), G > H, and F > H; and then rank them in the resulting order of preference, i.e. F, G, H. This much gives us an *ordinal* utility ranking, in which we can ascertain the agent's preferences without regard to the numerical values that may be assigned to them. Such values may be arbitrary so long as they descend from F to H, or absent altogether. So although such a ranking certainly may be combined with that given by other agents (voting would be the obvious example), its values cannot be interpersonally compared. That is, they cannot be manipulated arithmetically relative to others in order to derive a social utility function.

By contrast, a *cardinal* utility ranking of the agent's preferences would produce a fixed proportion of intervals on a preference scale that could be calibrated by the functions *n, n+1* for any *n*. Following Edgeworth,[4] assume that the agent can distinguish only a finite number of utility levels; and is indifferent between alternatives at the same level. Then a cardinal measure of the utility of different alternatives to an agent would be the number of levels on that agent's utility scale that separate them. But cardinality alone does not ensure interpersonal comparability. To make interpersonal comparisons, we

---

[2] John Rawls stated but did not elaborate this thesis while teaching a class in Social and Political Philosophy in 1974. In this section I attempt that elaboration.

[3] See John Broome, "Utilitarianism and Expected Utility," *The Journal of Philosophy LXXXIV*, 8 (August 1987), 405-422 for a good discussion.

[4] Francis Ysidro Edgeworth, *Mathematical Psychics and Other Essays* (San Diego: James and Gordon, 1995), pp. 46-50, Appendix III ("On Hedonimetry"). Edgeworth derives these assumptions from Wilhelm Wundt, *Principles of Physiological Psychology*, trans. E. B. Titchener (New York: Macmillan, 1904).

then would have to further assume that the intervals between levels is the same for all agents, and then compare the different cardinalities assigned to each given option by each agent.

It is this last assumption that generates the problem. Why should we assume that the intervals between levels of utility are the same for all agents, when the perceived intervals between seconds and minutes and hours differ so radically from one agent to the next depending on age, circumstance, and neurochemistry? Different individuals might have different feelings of different qualitative intensity about different alternatives at the same level (for example, apples and oranges). Or different individuals might feel differently about the same increase or decrease in utility level, of the sort that might distinguish the response of an emerging novelist from that of John Updike to a single bad review following a succession of favorable ones. To take another example: if an hour can drag by for an eighteen-year-old but rush by for a sixty-year-old, the difference between liking an orange at five units and liking orange juice at one unit more similarly might be vast for the former and negligible for the latter. For these reasons establishing even a subjective cardinal utility function bodes trouble enough. The prospects of establishing an interpersonal one look even gloomier.

1.2. The Von Neumann-Morgenstern Cardinal Measure

The Von Neumann-Morgenstern (henceforth the vN-M) cardinal measure for choices under risk presupposes the behavioral interpretation of (U),[5] and amplifies and clarifies a method first developed by Ramsey. Designed to determine the expected utility of some gamble x for an agent, it is based on empirical observation of that agent's choices among gambles. It thereby answers the question, how willing is that agent to gamble on satisfying some preference given the weights and probabilities he assigns to the outcomes of the actions open to him? The answer – the total value V of that preference – is the sum of the weights of each of the possible outcomes $a_1, a_2, \ldots a_n$ times their respective probabilities $p_1, p_2, \ldots p_n$, i.e.

$$V = a_1p_1 + a_2p_2 + \ldots + a_np_n.$$

The vN-M method derives from three preference axioms stipulating conditions on the relation '>' for the set S of all probability distributions or gambles on a set of outcomes:

[5] Although it should be noted that Morgenstern is critical of revealed preference theory. See Oskar Morgenstern, "Thirteen Critical Points in Contemporary Economic Theory: An Interpretation," *Journal of Economic Literature 10* (1972), 1163-1189.

(a) *Weak Ordering:* '>' is complete and transitive, i.e.
(i) for any F, G in S either F > G or G > F; and
(ii) if F > G and G>H then F>H;
(b) *Independence*: if F > G and 0 < p < 1 then
F(p) + H$(1 - p)$ > G(p) + H$(1 - p)$ for any H in S;
(c) *Continuity*: if F > G and G > H then F(p) + H$(1 - p)$ > G and
G > F(p') + H$(1 - p')$ for some real numbers p and p' between zero and 1.

The method consists, roughly, in two steps. The first is to arbitrarily assign weights of 1 and 0 to two alternatives F and H. The second is to then determine the expected utility to the agent of the third alternative G by finding the lottery of F and H that is indifferent to G, under the assumption that the agent's preference ordering of F, G and H satisfies constraints (a) – (c). So we first assume the agent's ranking of F, G and H:

F > G, G > H, F > H.

Then we assign a weight of 1 to F and 0 to H, and try to find the expected utility of G for the agent. We do this by constructing a choice situation for him: Either he will definitely get G – call this the *certain option*; or else he will get either F with a subjective probability of (p) or H with a probability of $(1-p)$ – call this the *lottery option*. Otherwise represented, the choice situation for that agent is

G or [F(p) or H$(1 - p)$].

For example, suppose p = .95 for Mabel. Then $(1 - p)$ = .05. So Mabel's choice is between G for certain on the one hand, and a 95% probability of F and a 5% probability of H on the other. Since Mabel has ranked F highest and the probability of F is awfully close to a sure thing, Mabel should choose the lottery option. On the other hand, what if p = .05 and $(1 - p)$, correspondingly, = .95? In this case Mabel's choice is between G for certain on the one hand, and, on the other, only a 5% probability of Mabel's highest-ranked option – F – and a 95% probability of Mabel's lowest-ranked option – H. In this case it would be more prudent for Mabel to choose the certain option – G. In general, as p changes from 1 to 0, Mabel's preference for the lottery option becomes a preference for the certain option because as her probability of getting her highest-ranked option decreases, the attractiveness to her of the certain option increases.

Now suppose there is a point at which Mabel is *indifferent* between the certain and the lottery options, i.e.

$$G \approx [F(p) \text{ or } H(1 - p)].$$

So, for example, suppose that $p = .66$. Then

$$G \text{ v } [F(.66) \text{ or } H(1 - .66)].$$

Remember that F = 1 and H = 0. So

$$G \approx [1(.66) \text{ or } 0(.33)],$$

i.e.

$$G \approx (.66).$$

So the expected utility of G for Mabel is .66, or 2/3. If F > H and $G = F(p) + H(1 - p)$, then F > G > H for $0 < p < 1$. By assigning a weight of 1 to Mabel's highest-ranked option F and 0 to her lowest-ranked option H, the vN-M method enables us to assign a cardinal value to any option in between them – without Edgeworth's assumption that intervals along individual utility scales are equal for all agents.

1.3. Interpersonal Cardinality

The vN-M method enables the construction of a cardinal utility scale. But it does not enable the construction of an *interpersonal* utility scale. For there is no independent way of locating our respective 1s and 0s relative to one another, nor of determining objectively the unit value of any interval between them. For example, suppose you observe my choice behavior over a broad array of pairwise comparisons. Suppose you then ask me to assign an aggregate numerical value, in increasing order of desirability, to each option chosen, such that each such value is the product of that option's weight for me, multiplied by the probability I assign to achieving it. Assume for now that my choice behavior is transitive and my options complete (these assumptions will be discussed in greater detail later). Suppose I give watching "The Simpsons" a 3, reading Trollope Sr. a 6, and listening to Mozart's "Jupiter" Symphony a 9. Then you could conclude that listening to Mozart's "Jupiter" Symphony was more satisfying to me than the other two, and that reading Trollope Sr. was more satisfying to me than watching "The Simpsons" by certain measurable proportional intervals. This would give us a cardinal ranking of my preferences. But it would not enable us to compare my utility rankings with yours, such that the members of both sets might be assigned objective rankings relative to one another on a single scale.

Why not? Because there is no way to rule out the possibility, nor to confirm it, that you express *the very same responses* towards "The Simpsons,"

Anthony Trollope , and Mozart's "Jupiter" Symphony, by assigning them the aggregate values of 1, 2, and 3 respectively that I expressed by assigning them the aggregate values of 3, 6 and 9 respectively:

Choice Options	*Me*	*You*
"The Simpsons"	3	1
anything by Trollope Sr.	6	2
Mozart's "Jupiter"	9	3

I.e. we both might agree that the "Jupiter" was better than Trollope Sr. and that Trollope Sr. was better than "The Simpsons." This much would yield agreement in our *ordinal* comparisons of interpersonal utility. Further, we both might agree on the proportional relations among them. And we both *in fact* might assign each option the same objective quantity of aggregate value. Despite all that, we might still diverge in our *numerical* representations of those quantities when assigned to a place on our respective scales, in how much quantity our respective scales could accommodate, and in how finely calibrated each such scale was. This point must be kept in mind when we address the quantitative representation of preference rankings in Sections 2.1 – 2.2 below and in Volume II, Chapter III.7. A cardinal utility scale without a viable method for making interpersonal comparisons among such scales does not suffice for the quantitative representation of preferences.

In this case, our respective utility functions might be distinguished by saying that I (or, of course, you) had a wider range of response capabilities, discriminated more finely, had higher expectations of satisfaction, and so on. But there would be no objective basis, independent of the incommensurable numerical rankings of each, for establishing these objective differences. Since there would be no way of ascertaining whether your 1 referred to the same objective quantity as my 1, there would be similarly no way of ascertaining whether or not your .5 might be the same as or different from my .1.

This could easily lead to mischief of a sort that merely ordinal interpersonal comparisons do not avoid. When ranked objectively on a single, numerical scale, my least preferred alternative would be assigned the same aggregate value as your most preferred one:

-10
-9 – Mozart's "Jupiter" (me)
-8
-7
-6 – Trollope Sr. (me)
-5
-4
-3 – "The Simpsons" (me), Mozart's "Jupiter" (you)
-2 – Trollope Sr. (you)
-1 – "The Simpsons" (you)

Some third party, appointed to collate a social utility function from our respective preferences, might well reason that I didn't seem to mind Trollope Sr. all that much, whereas you didn't seem to care for Mozart's "Jupiter" all that much more than Trollope Sr. On that basis we both might be sentenced to a steady diet of Trollope Sr. – even though both of us *in fact* most preferred Mozart's "Jupiter," and to the same degree. The subjective degree of our respective preferences for each of these alternatives cannot be ascertained despite their aggregate numerical values. So although each of us can assign cardinalities to the options presented, my preferences cannot be given a verifiably accurate cardinal ranking on the same scale as yours.

The difficulty is nevertheless not a special case of a private language, first-/third-person asymmetry, or other minds problem. No such calibration could be objective in the sense utility theory requires, even if the term "utility" had some fixed reference to a qualitatively identifiable inner state, and even if that state were empirically observable and individually quantifiable by means of some overt behavioral manifestation. To see this, suppose there were some sort of natural physiological barometer that all human agents had, such as a pale pink "utility mole" in the middle of our foreheads that turned bluer as one felt more overall satisfaction. Suppose further that my utility mole turned bright cobalt blue when I received $500.00, whereas yours attained that hue only upon receiving $500,000.00. What would that demonstrate? Surely not that I was objectively more satisfied overall with my $500.00 than you were with your $500.00. My satisfaction with my $500.00 might still be less objective satisfaction quantitatively than your dissatisfaction with yours, even though my utility mole is bluer. And surely not that I was just as objectively satisfied overall with my $500.00 as you were with your $500,000.00. My satisfaction with my $500.00 might still be far less objective satisfaction quantitatively than yours with your $500,000.00, even though our utility moles were the same shade of blue.

Thus the problem about making interpersonal comparisons of objective utility does not disappear by conjuring a solution to the problem of other

minds. The faith that such a solution would make the problem of interpersonal comparisons go away is based on conceiving of the third-personal, intersubjectively verifiable mechanism calibrating individual utility as though it were a thermometer measuring a temperature. However, if maximizing utility were relevantly similar to running a fever, then the problem of interpersonal comparisons would not arise. It is because it is not that a thermometer does not help to measure it. The utility mole example shows that the problem of interpersonal comparisons is caused, not by the existence of other minds, but rather by the existence of different *subjects*. It is not the inaccessibility of a subject's inner states, but rather his subjectivity itself that presents the obstacle to interpersonal comparisons of utility.[6] However, that interpersonal comparisons are in theory impossible to make does not imply that there is no objective fact of the matter about whether two individuals are equally satisfied or not.

So interpersonal comparisons of utility are in theory impossible because it is impossible to compare utilities among subjects, whether or not those subjects' mental states are interpersonally accessible. If individual cardinal utility rankings cannot be interpersonally compared, interpersonal comparisons of utility cannot be made. Hence neither can meaningful interpersonal *ascriptions* of utility. There can be in theory no shared criterion of cardinal utility-maximization that would enable me to confirm the actual degree of utility-maximization you reveal in your behavior from the degree of utility-maximization I ascribe to your behavior. If there can be no shared interpersonal criterion of cardinal utility-maximization, then there can be no nonvacuous but universal criterion of cardinal utility-maximization that applies to the evaluation of each and every case, even under conditions of risk or uncertainty. So long as (U) is claimed to have universal application, such ascriptions will be vacuous whether "utility" is interpreted as a mental state or not.

1.4. Allais on Psychological Value

Maurice Allais vehemently opposes this conclusion. He argues that cardinal utility, which he calls *psychological value*, "is fundamental to the theory of random choice [i.e. decisions under conditions of risk] (45)" and "an undisputable [*sic*] reality(8)."[7] He also believes an index of cardinal utility can

[6]Thus I disagree with Allan Gibbard, who conceives the problem of making interpersonal comparisons as a special case of the problem of knowing other minds. See his "Interpersonal Comparisons: Preference, Good, and the Intrinsic Reward of a Life," in *Foundations of Social Choice Theory,* Edited by Jon Elster and Aanund Hylland (New York: Cambridge University Press, 1989), 165-193.

[7] "Fondements d'une Théorie Positive des Choix Comportant un Risque et Critique des Postulats et Axiomes de L'Ecole Americaine,"Memoir III of *Econometrie XL* (1953), 257-

be ascertained only through "introspective observation of either psychologically equivalent increments or minimum perceptible thresholds (35)" as a basis for answers to "appropriate questions following processes similar to those used by psycho-physiologists such as Fechner and Weber (8)."[8] Through such directed questioning, psychological value, according to Allais, can be operationally defined and therefore measured in accordance with the Fechner-Weber laws concerning differential thresholds, or JNDs (just noticeable differences) (33, 43-44, 46). Allais' position in essence returns us to the phenomenological interpretation of (U) addressed in Section 4.1.

Let us put aside the question of whether Allais' cardinal utility index is being derived from or presupposed by this procedure, in order to look more closely at the Fechner-Weber laws which Allais claims provide the standards and methods for measuring psychological value. The Fechner-Weber laws, formulated in the mid-nineteenth century, express proportional relations between an external sensory stimulus, such as a physical weight or tone, and the subjective sensation it causes. These relations are established by experimentally measuring, for example, the increase in physical weight a subject lifts that is necessary for her to distinguish it as heavier than the first; or the distance apart on a region of skin two points must be in order for the subject to sense them as two. If R is a standard stimulus amount, ΔR is the just noticeable stimulus difference (i.e. the subjective sensation), and K is a constant, then according to Fechner,

$$(1)\ \Delta R/R = K.$$

Next we assign a value of 0 to the *absolute threshold stimulus*, i.e. that at which the subject's sensation of the stimulus is just about to appear, and assume that JNDs are equal within a given sensory modality. Then arithmetically increasing ΔR by steps of one multiplies the value of R by a constant ratio, and it becomes possible to measure and compute magnitudes of sensation relative to the stimuli that cause it. Because the subjective sensation increases

---

332 (Colloques Internationaux du Centre National de la Recherche Scientifique, Paris), translated as "Foundations of a Positive Theory of Choice Involving Risk and a Criticism of the Postulates and Axioms of the American School," in Maurice Allais and Ole Hagen, Eds. *Expected Utility and the Allais Paradox* (Dordrecht, Holland: D. Reidel, 1979), 27-146. Henceforth references to this article will be paginated in the text.

[8] To this Marschak rightly points out that answers are empirically observable behavior, too. He says, "[T]he 'introspective' comparison of 'satisfaction increments' by the subject provides the experimenter with words not with observed choices; … I suppose we are more interested in predicting actions than words; and such predictions are probably better based on recorded actions than recorded words." J. Marschak, "Utilities, Values, and Decision Makers," in Allais and Hagen, *ibid.* Note 7, 168.

arithmetically by a constant difference while the external stimulus increases geometrically by a constant multiple, the final form of Fechner's law is

(2) S = K log R,

where S is the magnitude of the sensation, K is a constant, and R is the magnitude of the stimulus.[9]

But whether JNDs can be assumed to be equal, even within a given sensory modality, is – as before – open to question. On what grounds can a subject (let alone the observer) conclude that the perceived difference between weightlessness and noticeable physical weight is equal to the difference between noticeable weight and just noticeably heavier weight? Or between the latter and weight just noticeably heavier than that? In order to justify this assumption, there would have to be some way of measuring, not only the material increases in physical weight of the respective stimuli relative to the sensation of increase reported by the subject; but also the sensed intervals between each of these increases. It is hard to imagine how these sensed intervals themselves could be measured; or why, in the absence of such measurement, they should be assumed to be equal.

However, there are more serious problems with Allais' reliance on this theory. The Fechner-Weber laws have been long since overtaken by more recent developments in experimental psychology. For example, it transpires that their ratios differ according to the sensory modality. Also, they hold only for the middle range of stimuli and not for extremes at either end of the spectrum. Further, what counts as an extreme differs from one subject to another, as do absolute stimuli thresholds.[10] But putting aside even these doubts, what relevance can the Fechner-Weber laws have for measuring the cardinal utility of various complex outcomes of action? Are we supposed to count the number of sensory modalities involved in each projected outcome and sum their respective ratios? Or multiply them? What if their sensory modalities have nothing to do with their psychological value for a particular agent? What if a psychologically valuable outcome engages none of the agent's sensory modalities, but rather her intellectual or aesthetic discrimination? What if the gambles an agent confronts involve, not simply various sums of money, as Allais supposes, but more complex but pedestrian intangibles such as status, intellectual stimulation, or self-worth, none of which can be realistically calibrated in terms of monetary gain?

---

[9] G. T. Fechner, *Elements of Psychophysics*, Vol. I, Trans. H. E. Adler, Ed. E. G. Boring and D. Howes (New York: Holt, Rinehart and Winston, 1966).

[10] These more recent developments are summarized in Robert Watson, *The Great Psychologists: From Aristotle to Freud*, Second Edition (New York: J. B. Lippincott Co., 1968), 232-239.

Allais depends on these laws to ground his claim that for a given field of choice defined by an ordinal preference function

(3) S = S(A, B, … C),

such that A, B, … C are quantifiable commodity bundles, "it is possible to specify a cardinal preference function

(4) $\underline{S}$ = $\underline{S}$(A, B, … C)

[in which $\underline{S}$ is the corresponding expected utility function], determined up to a linear transformation, such that for two increments of utility that are deemed equivalent, the variation $\Delta\underline{S}$ will have the same value (43)." But he does not explain how or why two "increments of utility" are to be "deemed equivalent" any more than Fechner did. He does argue as follows:

> Hardly anybody would fail to reply 'yes' *without any hesitation* if asked 'would you prefer to inherit $100 million rather than $10,000 more strongly than you would prefer to inherit $10,000 rather than $1,000?' The absence of hesitation demonstrates without any doubt that the notion of equivalent psychological increments indeed corresponds to a psychological reality (46).

That is, Allais reasons that because we would unhesitatingly most prefer a larger amount of money that is ten thousand times more than what we would least prefer, to a smaller amount that is only ten times more than what we would least prefer, we must have a concept of equivalent psychological increments. But in order to infer from our ranking of the proffered monetary sums that we have a concept of equivalent psychological increments, this argument must presuppose that monetary value is equivalent to psychological value. Allais does not defend this presupposition. Instead he offers the suggestion, in a footnote, that "[i]t is worth recalling that the problem of determining cardinal utility is identical to the problem of determining the marginal utility of money (n. 17)." He also simply states in passing that to each monetary value there corresponds such a measurable psychological value, or cardinal utility (45-46). But if Allais' experiments are supposed to yield conclusions merely about agent responses to the probabilities of receiving certain sums of money, then he does not need the concept of psychological value; whereas if they are to have implications for a broader and more complex range of cases of the sort described above, then his equation of monetary with psychological values is inadequately defended.

On the basis of this equation, Allais then draws a further analogy between the comparison of (3) and (4) and a second one. Given a field of choice among risky outcomes defined by the function

(5) $S = S(g_1, g_2, \ldots, g_n, p_1, p_2, \ldots, p_n)$,

Allais proposes a function of cardinal preference

(6) $\underline{S} = \underline{S}(g_1, g_2, \ldots, g_n, p_1, p_2, \ldots, p_n)$,

"which," he claims, "in the field of the economic psychology of risk, is the equivalent of the Fechner-Weber functions expressing psychophysiological sensation as a function of excitation (44)." Note that Allais is claiming that formula (6) is the *equivalent*, in economic choice under conditions of risk, of Fechner's (1) and (2), above, in the psychology of physiological stimulation. He is not merely claiming that they are related, or analogous. Nor is he claiming merely that there is some correspondence between them. He is claiming that they are equivalent. However, he does not show in what the equivalence consists. What remains is, first, to explain the relevance of this seemingly anachronistic psychological theory to the construction of a cardinal utility scale. Second, the sense in which its laws are equivalent to (6) is also in need of clarification. Without this, a convincing case even for the existence of a single subject's cardinal utility scale, much less for the possible of comparing such scales interpersonally, has not been made.

1.5. The Allais Paradox

Allais grounds his rejection of the vN-M method in the view that an individual subject must be asked to introspect on her subjective, phenomenological response to alternative gambles with weights and subjective probabilities that differ radically in some cases and very minimally in others. He shows that the outcome of this experimental procedure generates real life counterexamples to the vN-M independence condition (E.2.(b), above), and therefore uncovers inconsistencies in rational choice that undermine its universal scope.

Now the vN-M axioms were not intended to be universal in scope. As we will shortly see, it is true that the assignment of huge weights, infinitesimal probabilities, or minute incremental differences to preference alternatives can in some cases produce counterexamples to the vN-M axioms. But Morgenstern rightly protests that the method is not designed to extend to such extreme cases.[11] In Section 1.6 following, however, I argue that there is in theory no way of ruling out such cases; and that their necessary inclusion sabotages the attempt to exclude cyclical rankings through the imposition of normative requirements such as transitivity, irreflexivity, or independence.

[11] See Oskar Morgenstern, "Some Reflections on Utility," Allais and Hagen, *op. cit.* Note 7, esp. 178.

Second, the force of Allais' experimental counterexample does not depend on his phenomenological interpretation of (U). The case presents an agent with two sets of compound gambles. In (1), the agent is presented with a choice between F and G, such that

(1) F = \$500,000 with a probability of 1 (the "sure thing"); and
G = ◊\$2,500,000 with a probability of .10, or
◊\$500,000 with a probability of .89, or
◊\$0 with a probability of .01.

In this case, the agent prefers F to G. In (2), the agent is presented with a choice between H and I, such that

(2) H = ◊\$500,000 with a probability of .11, or
◊\$0 with a probability of .89; and
I = ◊\$2,500,000 with a probability of .10, or
◊\$0 with a probability of .9.

Here the agent prefers I to H. So she prefers a certain tidy sum to an unlikely jackpot, but an equally unlikely jackpot to an only slightly less unlikely tidy sum. That is, she chooses the risk-averse option in case (1), but the riskier option in case (2). We can see the dilemma by parsing the payoffs in thousands as follows:[12]

(1)

	.89 [E_1]	**.01 [E_2]**	**.10 [E_3]**
F	\$500	500	500
G	\$500	0	2500

(2)

H	0	500	500
I	0	0	2500

By partitioning E_1 and (E_2 or E_3), we see that options F and H have identical payoffs under E_2 and E_3, as do options G and I. According to the independence axiom, if F is preferred to G, then (500,*p*; 500 1 – *p*) is preferred to (0, *p*; 2500, 1 – *p*). But then H should be preferred to I. Substituting 0 for \$500 under E_1 should make no difference. But Allais' experiment shows that most subjects prefer I to H because although the probability of winning is almost identical in both gambles – 11% in H, 10% in I, the payoff in I is much

[12] Here I am grateful to Ned McClennen.

larger. It shows that the behavior of a seemingly otherwise rational chooser may disconfirm independence when it is interpreted as descriptive, predictive and explanatory; and violate it when it is interpreted as normative. But Allais' experiment does not show that requiring subjective introspection from such a subject is needed to do this; nor, therefore, does it make a case for an alternative to the vN-M measure of cardinal utility.

Now Morgenstern, in refuting Allais' experimental results with his own, finds nothing objectionable in "educating" his experimental subjects as to the theory their behavior is predicted to fit, in order to insure the veracity of the prediction.[13] But in standard experimental procedure, coaching one's subjects in order to achieve the desired result predicted by the theory would violate prohibitions on tampering with the data. Morgenstern rejects this concern on the grounds that since his theory is normative, his student subjects are merely "correcting" their behavior in light of what they believe, as the result of his instruction, to be rational. However, a theory cannot function as both normative and empirically predictive for the same subjects under experimental conditions without invalidating the experiment as a legitimate test of the theory's explanatory power. Morgenstern then comments

> Naturally, it is assumed that the individuals are accessible intellectually whether it be physics or arithmetics [*sic*] or utility that is being explained to them.]In that sense there is a limitation since there are certainly persons for whom this is impossible.]Whether they then should be called "irrational" is a matter of taste (180).

It appears that those who fail to adjust their behavior according to his instruction must be stupid, and perhaps irrational. This would make a universalized version of the von Neumann-Morgenstern cardinal utility function both normatively *and at the same time* descriptively vacuous. This unhappy implication can be avoided by decisively rejecting the claim of universality.

1.6. Preference and Probability

We have seen that Allais' phenomenological interpretation of utility-maximization does not redress the insufficiencies of the vN-M measure for interpersonal comparisons, and therefore does not block the conclusion to the vacuity of (U). We have also seen that the Allais Paradox depends on the assignment of weights and probabilities of extreme magnitudes to preference alternatives; and that the vN-M cardinal utility measure excludes these as beyond the scope of the weak ordering, independence and continuity axioms. In this section I argue first, in 1.6.1, that Allais assignments must be admissable in an agent's preference ranking – thereby generating the Allais Paradox in certain cases; and in 1.6.2 that the same considerations that

[13] Allais and Hagen, *op. cit*. Note 7, 180.

guarantee the admissibility of Allais assignments to preference rankings also generate cyclicities that structurally violate *all* of the von Neumann-Morgenstern axioms, and by implication several others. From this conclusion I infer that imposing upon an agent's ranking any such normative axioms of choice fails to exclude cyclicities; and that these structural inconsistencies hasten the conclusion to (U)'s vacuity. Sections 2 and 3 below detail the close connections between cyclicity and vacuity; and Volume II, Chapter III both exposes their logical relationship (essentially, the relationship between self-contradiction and tautology); and also proposes a way to avoid both.

1.6.1. Aggregate Value and the Sorites Paradox

The argument is grounded in an implicit tension between the concepts of preference and probability. The concept of a consistent preference ordering implicitly presupposes observation of the all-or-nothing law of noncontradiction as understood in predicate logic – the principle of bivalence, as vagueness theorists call it; whereas weight and probability assignments to projected outcomes presuppose a linear mathematical progression. Some simple examples show how these two sets of presuppositions may conflict.

Consider first a series of cases in which an agent must choose among three simple, weighted tidy sum-alternatives F, G, and H, such that in each case the three consistently weighted alternatives receive different probability assignments relative to the others. An infinitesimal probability assignment or incremental difference in probability assignment may reduce a consistently weighted alternative's aggregate value so much that it becomes a non-alternative for that agent. The gradient metamorphosis of an alternative into a non-alternative generates a sorites paradox that thereby creates an inconsistency in the set of alternatives available to him, and therefore in the choices he makes as a consequence.

For example, suppose Percy must choose among F = \$100, G = \$90, and H = \$80; and that Percy chooses in accordance with (U), such that this choice satisfies completeness and transitivity (vN-M's 1.2.(a) i. and ii., above). Then F > G, G > H, and F > H.[14] Now consider the following three sets of probability assignments to F, G, and H, and their effect on the aggregate value to Percy of each alternative. In the first, we assume that each alternative is a sure thing:

[14] We can ignore the distinction between monetary worth and its intrinsic or marginal value for purposes of this example.

(1)

	weight	**x** *probability*	**=** *aggregate value*
F	$100	1	100
G	$90	1	90
H	$80	1	80

Since the alternatives do not constitute a gamble, the assignment of a probability of 1 to each alternative does not change its aggregate value to Percy – nor, therefore, its place in Percy's preference ordering: He still prefers F to G and G to H. By contrast, the second set of probability assignments does present Percy with a gamble:

(2)

	weight	**x** *probability*	**=** *aggregate value*
F	$100	.2	◊20
G	$90	.000000000001	◊.000000000009
H	$80	.799999	◊63.99992

In this case, the probability of achieving each alternative revises the aggregate value of each to Percy, and so changes his preference ordering. Although he still prefers F to G, he now prefers H to F because H is close enough in weight to F that the higher probability of getting H considerably increases its aggregate value. So Percy prefers H to F and F to G, and therefore H to G. This ranking is neither irrational nor remarkable in its degree of risk aversion. The third set of probability assignments has a different effect:

(3)

	weight	**x** *probability*	**=** *aggregate value*
F	$100	.2	◊20
G	$90	0	□0
H	$80	.8	◊64

It would seem that under this probability assignment, G is ruled out as a preference alternative because the probability of its occurrence – and so its aggregate value – is zero, despite its dollar weight. (Similarly, if there is no chance whatsoever that Rupert Murdoch will give me a billion dollars, it makes no sense for me to claim that I prefer a modest but steady paycheck.) So G has no place in Percy's preference ordering. Although Percy does prefer H to F, it is therefore not the case that Percy prefers F to G, and so not the case

that he prefers H to G. This would seem to violate transitivity (1.2.(a.ii), above).

But if G in (3) is not one of Percy's live options, how can it cause the violation of a condition imposed on them? Easy: by being assigned a weight, which assures it a location in Percy's ordinal preference ranking; but zero computable probability, which eliminates its positive aggregate value. Like the weak ordering axiom more generally, the requirement of transitivity is not contingent on the particular numerical probability assigned to each alternative. Transitivity is required to hold *whatever* those probabilities may be.

Yet the difference between G's aggregate value in case (2) and its aggregate value in case (3) is very, very small. If G is not a live option in case (3), is it one in case (2)? And how much would we have to pump up the weight of G in case (2) in order to convince Percy to see it as such? Von Neumann-Morgenstern, in rejecting Allais assignments, might plausibly contend that the likelihood of G is sufficiently negligible in both cases that it is not a live option in either; and that we should simply discount G's pro forma status as a ranked alternative in case (2) as well as (3). Call this the *Contra-G Argument*.[15] Then since G is not a preference alternative, it is no truer in case (2) than it is in case (3) that Percy prefers F to G; nor, in either, that he prefers H to G.

On the other hand, Allais might argue with equal plausibility that case (2) leaves room for some minimal expectation of G, however unrealistic, whereas case (3) leaves none; and that therefore, G's status as a preference alternative in case (2) must be honored. Of course an expectation this minimal is not merely unrealistic but negligible. But so long as there is a probability, however minute, that G, a correspondingly minute aggregate value can be attached to it. So G remains a live option despite its infinitesimal probability. Call this the *Pro-G Argument*. The Pro-G Argument implies that Percy does after all prefer F to G and so H to G; and therefore satisfies transitivity.

In either case, the extent to which G qualifies as a live option or not is a matter of degree, dependent on the agent's – and our – evaluation not of its probability, but of the *significance* of its resulting aggregate value. If the likelihood of some option O is small enough, it will be imperceptible even to the agent whose option it is. O's aggregate value will decline toward zero accordingly, to a point at which it, too, is imperceptible to the agent, regardless of its weight. As O's aggregate value declines to imperceptibility, the significance of its aggregate value correspondingly declines towards negligibility.

---

[15] Samuel Gorovitz's "The Saint Petersburg Puzzle" (Allais and Hagen, *op. cit.* Note 7, esp. 265-268) might be understood as an extended Contra-G Argument.

Notice that we could run the same argument on weight, keeping probability fixed and certain:

(4)

	weight	**x** *probability*	**=** *aggregate value*
F	2	1	2
G	.000000000001	1	.000000000001
H	6	1	6

(5)

	weight	**x** *probability*	**=** *aggregate value*
F	2	1	2
G	0	1	0
H	6	1	6

Is G in (5) a live option or not? Even if Inez is one hundred percent certain that G, which is worthless to her, will obtain if she so chooses, its utter lack of value in her eyes does not merely reduce it to a lowest-ranked option. Its zero aggregate value decisively *rules it out as an option* for her. The difference between a state of affairs that is a lowest-ranked option and one that is not an option at all is that the latter lacks any discernible qualities that might persuade the agent to choose it, even in the worst-case scenario.

If G's zero aggregate value rules it out as one of Inez's options in (5), then it is hard to see why she should count it among her options in (4) either, given its negligible aggregate value – and so indiscernible redeeming qualities – there. In both cases, aggregate value exists along an asymptotic linear continuum that shades off imperceptibly to zero at its lower limit. Here, too, if the agent's weighting of option O is small enough, it will be imperceptible. O's aggregate value will similarly decline toward zero and hence imperceptibility regardless of its probability. As O's aggregate value declines to imperceptibility, the significance of its aggregate value again correspondingly declines towards negligibility.

Both the contra-G and the pro-G arguments, both about probability and about weight, generate sorites paradoxes:[16]

[16] For a lucid analysis of the sorites paradox, see Stephen Schiffer, *The Things We Mean* (New York: Oxford University Press, 2003), Chapter 9: "Vagueness and Indeterminacy."

(a) THE CONTRA-G [VON NEUMANN-MORGENSTERN] SORITES PARADOX

(i) A weighted preference alternative of zero probability has no aggregate value.

(ii) A weighted preference alternative of (0 + .000000000001) probability has no aggregate value.

(iii) A weighted preference alternative of (0 + .000000000002) probability has no aggregate value.

.
.
.

(iv) ∴ A weighted preference alternative of .2 probability has no aggregate value.

(v) ∴ No weighted preference alternative of any probability has aggregate value.

(b) THE PRO-G [ALLAIS] SORITES PARADOX

(i′) A weighted preference alternative of .2 probability has aggregate value.

(ii′) A weighted preference alternative of (.2 – .000000000001) probability has aggregate value.

(iii′) A weighted preference alternative of (.2 – .000000000002) probability has aggregate value.

.
.
.

(iv′) ∴ A weighted preference alternative of 0 probability has aggregate value.

(v′) ∴ Every weighted preference alternative has aggregate value regardless of its probability.

In both (a) and (b), the conclusions (iv)-(v) and (iv′)-(v′) respectively are false. In both (a) and (b), the first three members of the infinite series of premises [(i), (ii), (iii) …] and [(i′), (ii′), (iii′), …] respectively are true. But in the infinite series of premises neither of (a) nor of (b) is there a cut-off point that decisively marks the distinction between having aggregate value and having none; nor, therefore, a viable criterion for excluding Allais assignments from the corresponding preference rankings. The concept of aggregate value is vague.

Stephen Schiffer proposes to solve the sorites paradox by narrowing the scope of the principle of bivalence in classical logic – i.e. that every proposition is either true or false. For Schiffer, whether bivalence applies to non-tautological propositions containing vague concepts – i.e. those without

determinate cut-offs between truth and falsity – is indeterminate.[17] Although Schiffer would disagree,[18] we in fact often say that such concepts – for example, baldness or wealth – hold true to different degrees; that propositions containing them are more or less true. But the concept of aggregate value – like the concepts of weight, probability or degree itself, and unlike the concepts of baldness or wealth – already is precisely that sort of formal quantitative and gradient concept of degree to which the principle of bivalence typically yields in propositions containing vague concepts. Yet in order successfully to apply that concept of degree itself in (a) and (b), the principle of bivalence must be invoked: it either is or is not true that whether G in (2) has aggregate value or not is a matter of degree. If we answer that

> ^whether G in (2) has aggregate value or not is a matter of degree^

is itself a matter of degree, the unavoidable reply is that it either is or is not true that

> ^^whether G in (2) has aggregate value or not is a matter of degree^ is itself a matter of degree^

and so on. (a) and (b) demonstrate that even concepts of degrees of shading admit of degrees of shading; and so that abandoning bivalence exacerbates rather than dissolves the sorites paradox, by generating an infinite regress of concepts of degrees of shading.

A resolution to the question of whether G in (2) has aggregate value or not requires some means of differentiating between having and not having a particular degree of aggregate value. This is the principle of bivalence. To this question, Allais would answer affirmatively for all probability assignments to G excluding (3). Von Neumann-Morgenstern would beg to differ; and perhaps might marshal a large sample of experimental subjects to settle the matter statistically by voting. But if whether any preference alternative has aggregate value is, as (a) and (b) imply, a matter of degree, which itself is a matter of degree, etc., then no such means of differentiating between having and not having aggregate value can be found. Then whether any preference alternative has aggregate value or not is indeterminate; and the above argument applies to the concept of indeterminacy as well. Either it, too, is subject to the principle of bivalence; or else it is vague and so generates an infinite regress of degrees of indeterminacy.

---

[17] Schiffer argues that the paradox can be generated for any vague concept whatsoever (*ibid*., 178, n.1).

[18] See Schiffer's critique of degree-theoretic notions of truth at *ibid*., 191-194.

A sorities paradox can be generated for any judgment involving a quantitative concept, and *a fortiori* for any assignment of weights and probabilities to preference alternatives, whether they fall within the normal or the infinitesimal range. Therefore any Allais assignment must fall within the series of premises [(i), (ii), (iii), … (iv)-(v)] in (a), or [(i′), (ii′), (iii′), …(iv′)-(v′)] in (b). So it is hard to imagine on what special grounds, or according to what special criterion, the extreme assignments on which the Allais Paradox relies can be excluded from an agent's preference ranking. On the contrary: other things equal, certain preference alternatives (excluding death and taxes) may enter the outermost edges of an agent's field of choice with probability assignments that decrease in magnitude with their temporal distance from the agent, and, so long as their weights do not decrease to zero, increase in aggregate value as the agent moves forward in time to meet them. My preferences for having an apple for breakfast and residing in Berlin exactly twenty years from today would be of this kind. Alternatives that enter the field at the most temporally remote outer edges may well bear, at that point, the infinitesimal probability assignments on which Allais-style paradoxes thrive. They may remain preference alternatives nevertheless.

1.6.2. Sorites, Cyclicity and the vN-M Axioms

I have just argued that Allais assignments must be admissable in an agent's preference ranking – thereby empirically disconfirming the independence axiom via the Allais Paradox. But the same Pro-G or Contra-G arguments that generate sorities paradoxes also generate cyclical rankings that structurally violate all of the von Neumann-Morgenstern axioms, and indeed several others.

We have seen above that, by contrast with the vagueness built into the quantitative gradience of weight and probability assignments, whether or not Percy's ordinal ranking of the options available to him is consistent or not is (at least on the face of it) not a matter of degree but rather of classical logic. If G *is* among Percy's live options at a particular moment, then at that moment it either is or is not the case that Percy prefers H to F, F to G, and so H to G. However, that a sorites paradox can be generated in either case results from the fact that G's negligible aggregate value may figure in an argument on either side of this disjunction. G's aggregate value in case 1.6.1.(2) favors the Pro-G argument to the same extent that it favors the Contra-G argument. G's aggregate value would need to increase considerably in degree in order to shift the balance decisively in favor of the former over the latter, or else decrease to 0 in order to shift it in the opposite direction. In the absence of any such incremental increase or decrease, it both is and is not the case that Percy prefers H to F, F to G and so H to G.

From these unresolvable ruminations on case 1.6.1.(2) we can now easily generate formal violations of the other condition of the vN-M weak ordering

axiom (1.2.(a.i), above) that are recognizable using some of the conventional transformations of classical logic:

(1) F>G or G>F	Completeness
(2) not-F>G	the Contra-G Argument
(3) ∴ G>F	(1), (2)
(4) F>G	the Pro-G Argument
(5) ∴ F>G and G>F	(4), (3)
(6) not-F>G or not G>F	(2)
(7) not-G>F	(6), (4)
(8) F>G and not-F>G and G>F and not-G>F	

Both the pairwise cyclical ordering expressed in (5) and the straightforward logical contradiction expressed in (8) are structural rather than empirical or experimental. Therefore they are not susceptible to Fishburn's analysis of experimental violations of the weak ordering axiom either as "behavior [that] reflect[s] the individual's indecision about which of F and G he prefers" or as a case in which F and G "are 'close enough' that it is not worth his effort to be careful about which he chooses, [in which case] it seems reasonable in an operational sense to say that he is indifferent between F and G."[19] Rather, the individual – and we – are in a state of permanent and necessary indecision about how to classify G in case 1.6.1.(2). I discuss the distinction between such indecision and the indifference relation in Volume II, Chapter III.6.2. We cannot definitively answer the question as to whether G is a live option or not because this question requires a yes-or-no answer, whereas the extent to which it is or not is a matter of degree – which itself is a matter of degree, and so on. Under such circumstances, yes-and-no is the best answer we can give.

There is a deeper moral to this story. Allais' experimental result shows, like Tversky and Kahnemann's, among others, that empirical choice behavior may violate what we intuitively view as rational standards of consistent choice. A familiar and understandable response to such counterexamples is to try to protect consistency by imposing further conditions – transitivity, irreflexivity, independence, substitutability, continuity, etc. – on rational choice in order to rule out such counterexamples as instances of irrationality. But it is easily seen that the above argument need be only slightly reconfigured to generate similar structural cyclicities and contradictions for any such condition (I leave this as an exercise to the reader). Thus the same

[19] Peter C. Fishburn, "On the Nature of Expected Utility," in Allais and Hagen, *op. cit.* Note 7, 245.

rationality considerations that motivate the imposition of the conditions also generate the structural inconsistencies that undermine them.

This suggests that imposing such restrictive conditions on preference rankings may not be the most efficient way to protect the consistency of those rankings, because *counterexamples to those conditions remain* ***logically*** *consistent with them.* Instead of protecting preference consistency, imposing these conditions merely restricts the scope of application of the theory to the narrowly normative, thereby diminishing its empirical applicability and inviting structural inconsistencies like the ones above. Now we have already seen that a theory can be both explanatory and normative if it explains the behavior of an ideal agent who sets a standard we are exhorted to emulate. However, a theory that can be explanatory when and only when it is normative is not doing the work a theory is supposed to do. In Volume II of this project I take up the challenge to protect the consistency of rational choice using classical logic as a resource rather than rejecting it as a threat.

2. The Ramsey-Savage Concept of a Simple Ordering

In "Truth and Probability," Ramsey demonstrates that imposing certain consistency conditions on an agent's choices among an unlimited set of alternatives yields the interpretation that she seeks to maximize utility in her behavior, i.e. (U).[20] This is the original idea on which the theory of revealed preference is based. Ramsey's consistency constraints on preference rankings are generally assumed to rescue the behavioral interpretation of the concept of utility from vacuity or logical inconsistency. There are two reasons why they do not. First, particular axioms in Ramsey's system that help define in what consistency consists overlook the intensionality of preference and so at best spell out a sub-logical conception of consistency that does not clearly apply to it. Second, these constraints presuppose a more primitive concept of utility-maximization, namely (U), that is even more vulnerable to the reproach of vacuity under the revealed preference interpretation than under earlier ones.

2.1. Ramsey's Value Axioms

Ramsey begins with the assumption that

> [1] we act in the way we think most likely to realize the objects of our desires, so that a person's actions are completely determined by his desires and opinions … [2] we seek things which we want, which may be our own or other people's pleasure, or anything else whatever, and our

[20] Frank P. Ramsey, "Truth and Probability," in *The Foundations of Mathematics and Other Logical Essays,* Ed. R. B. Braithwaite (London: Routledge and Kegan Paul, 1950), 157-198.

> actions are such as we think most likely to realize those goods.[21] … [3] our subject has certain beliefs about everything; then he will act so that what he believes to be the total consequences of his action will be the best possible.[22]

Ramsey assumes that sentences [1] and [2] in the above passage are equivalent to [3]. [1] and [2] more accurately summarize the belief-desire model of motivation already discussed. However, [3] does not describe a motivational model. Rather, it is a prediction of intentional maximizing behavior, and as such encapsulates the utility-maximizing model of rationality, i.e. (U). [3] is in fact the foundational theoretical assumption underlying the constraints Ramsey offers.

Ramsey's project is to create a subjective probability measure of the degrees of belief on which a chooser's intention to maximize is based. His value axioms make it possible to calibrate and compare the value of different total outcomes F, G, H, I, J, K to a chooser based on the numerical value intervals among them. The basic idea is to measure my degree of belief in a sentence *s* by asking how much between 0 and 1 I would be willing to bet on the truth of *s*; the von Neumann-Morgenstern cardinal measure discussed in Section 1.2 above is a refinement on this basic idea. According to it, I believe *s* to a degree of only .5 if I am indifferent between the following two alternatives:

> (i) the truth of *s* secures total outcome F and its falsity secures G; and
> (ii) the falsity of *s* secures F and its truth secures G.

That is, nothing of consequence for me turns on whether *s* is true or false. If *s*'s truth gives it a probabilistic value of 1 and its falsity gives it a probabilistic value of 0, then my indifference between (i) and (ii) amounts to an indifference between the following two cases:

> (i') (1, F) and (0, G); and
> (ii') (0, F) and (1, G);

i.e. my ranking of total outcomes F and G remains unaffected by the possible truth or falsity of *s*. Call *s* a *toss-up* belief.

The difference in value I assign to F relative to G is equal to the difference in value I assign H relative to I if, given that my ranking of F and G remains

---

[21] *Ibid*., 173. By contrast with Ramsey, I. M. D. Little *defines* consistent behavior to include the maximizing motivational assumption ("A Reformulation of the Theory of Consumer's Behavior," *Oxford Economic Papers I* (1949), 91, 97).

[22] Ramsey, *ibid*., 176.

unaffected by the possible truth or falsity of *s*, I am equally indifferent between the following two options:

(iii) the truth of *s* secures F and its falsity secures I; and
(iv) the truth of *s* secures G and its falsity secures H.

That is, if nothing of consequence for my ranking of F and I turns on the truth or falsity of *s*, then similarly nothing of consequence for my ranking of G and H does, either. And similarly, this amounts to indifference between the following two cases:

(iii') (1, F) and (0, I);
(iv') (1, G) and (0, H);

i.e. my ranking of F and I relative to G and H remains similarly unaffected. In this case the value intervals between F and G, and between H and I are the same.

Then define as a *value* any set of outcomes I prefer to a given outcome, such that if I prefer outcome F to G, then I prefer any outcome with the same value as F to any outcome with the same value as G. In this case the value of F to me is greater than the value of G, and I can be said to rank total outcomes F, G, … in an ordinal series. Also define as an *ethically neutral sentence* (or proposition) *s* one whose truth or falsity makes no difference to the equal value of two possible worlds identical in all other respects.

Ramsey's first axiom (A1) stipulates the existence of such an *s* believed to a degree of .5, i.e. of an ethically neutral toss-up belief. His second axiom,

(A2) if *s, t* are such sentences and the option F if *s*, I if not-*s* is equivalent to G if *s*, H if not-*s*, then
F if *t*, I if not-*t* is equivalent to G if *t*, H if not-*t*

replaces the indifference relation between (iii) and (iv) above with an equivalence relation, and derives from it similarly equal value intervals for total outcomes F, G, H and I with respect to a second ethically neutral toss-up belief *t*. This establishes that the equality of the value intervals between F and G and between H and I is independent of the content of the ethically neutral toss-up belief on the truth of which I am willing to bet either way for the same stakes; and so, by definition, the equivalence of the intervals FG and HI. From this equivalence plus the operations of transposition and distribution, Ramsey derives the equivalence of the ordinal rankings F>G and H>I and of the equalities F=G and H=I.

The consistency of these intervals is secured by Ramsey's third and fourth axioms:

(A3) If option A is equivalent to option B and B to C, then A to C;

call this the Transitivity of Equivalent Options Axiom. (A3) establishes the unit consistency among comparable outcomes of whatever the unit quantity by which all of them are compared.

(A4) If FG = HI, HI = JK, then FG = JK;

call this the Transitivity of Equal Intervals Axiom. (A4) establishes the interval consistency among the intervals of measurement relative to which each outcome, conceived as some multiple of their unit quantity, is calibrated. Together axioms (A2) – (A4) ensure that, independent of the particular numerical values assigned to alternative outcomes, all such comparable outcomes can be ranked relative to one another as multiples of some unit quantity along a cardinal scale that assigns some numerical value to each.

There are eight axioms in all. (A5) and (A6) stipulate unit quantity uniqueness, (A7) stipulates ordinal continuity, (A8) the ratio between two numbers *a* and *b* given integers *m*, *n* such that $na>b$ and $mb>a$. Together they demonstrate how a value a might be correlated with a real number u(F) such that the value interval between F and G can be represented by the numerical expression u(F) – (u)G. I focus here on the reasoning behind (A2)-(A4) just summarized.

2.2. Consistency and Intensionality

From the outset, Ramsey's exposition relies on intuitive and unexplicated notions of preference and ranking. Axioms (A1) – (A8) articulate these intuitive notions with formal precision. But they do not function as foundations that these intuitive notions presuppose. The relationship is rather the reverse: The intuitive notions of preference and ranking provide the foundations that the formal axioms presuppose, and on the basis of which those formal axioms are interpreted. That is, we need to assume and accept the intuitive notions in order to make sense of the formalization. This means that implicit from the very beginning in the conception of utility-maximization and partial belief Ramsey develops formally is a more primitive, clearly intentional proto-concept of utility-maximization as basic action – (U), in fact, as defined in Chapter III, Section 1 above – that gives the formalized axioms meaning. Because this more primitive conception of utility-maximization is implicitly intentional, Ramsey's value axioms, in so far as they are valid, are implicitly intentional as well. The point in Ramsey's exposition at which intensionality is abandoned for the flexibility, precision and objectivity of extensional notation is the point at which these axioms cease to pertain to preference as we ordinarily understand that concept.

Now we have just seen that in the move from (A1) to (A2), Ramsey replaces the indifference relation between (iii) and (iv) that defines equality of value intervals FG and HI relative to an ethically neutral toss-up belief *s* with an equivalence relation that enables him to derive similarly equal value intervals relative to a second ethically neutral toss-up belief *t* with which *s* is, as regards content and degree of belief, interchangeable. This move deserves further scrutiny. Indifference between (iii) and (iv) was stipulated to be a sufficient condition for the equality of the value intervals of ranked total outcomes FG and HI. However, indifference is an intensional relation between two complex objects of value, whereas equivalence is an extensional relation between two complex sentences or propositions. To say that I am *indifferent* between (iii) and (iv) is to say that *it does not matter to my preference ranking of F, G, H, and I* whether *s* is true or false. This is what enables us to infer the equality of value intervals FG and HI.

By contrast, to say that (iii) and (iv) are *equivalent* is to say that (iii) is a *necessary and sufficient condition* for (iv), i.e. that the truth of *s* secures F and its falsity secures I *if and only if* the truth of *s* secures G and its falsity secures H. From this it follows that the truth of *s* secures F if and only if it secures G; and that its falsity secures I if and only if it secures H. To see this, use the Boolean connectives under their conventional interpretation in classical logic, and assume for purposes of this argument that F, G, H, I, J, and K can be interpreted as symbolizing not outcomes but rather sentences or propositions describing outcomes. Then (iii) becomes

(iii″) $(s \rightarrow F) \,.\, (\sim s \rightarrow I)$

and (iv) becomes

(iv″) $(s \rightarrow G) \,.\, (\sim s \rightarrow H)$.

Then (A2) becomes

(A2′) $[(s \rightarrow F) \,.\, (\sim s \rightarrow I) \equiv (s \rightarrow G) \,.\, (\sim s \rightarrow H)]$
$\Rightarrow [(t \rightarrow F) \,.\, (\sim t \rightarrow I) \equiv (t \rightarrow G) \,.\, (\sim t \rightarrow H)]$.

The antecedent of (A2′) can be rewritten as

(A2′a) $[(s \rightarrow F) \equiv (s \rightarrow G)] \,.\, [(\sim s \rightarrow I) \equiv (\sim s \rightarrow H)]$.

This makes the truth of *s* a sufficient condition for F if and only if it is a sufficient condition for G, and its falsity a sufficient condition for H if and only if it is a sufficient condition for I. But it does not show that the difference in value between F and G is the same as the difference in value between H

and I. Because, unlike the indifference relation, the equivalence relation does not assign value to its terms (not even equal value), it cannot show this.

Then the consequent of (A2), here rewritten as

(A2'c) $[(t \rightarrow F) \equiv (t \rightarrow G)] \,.\, [(\sim t \rightarrow I) \equiv (\sim t \rightarrow H)]$

follows as a statement about the independence of the *connective relations* among F, G, H, and I from the content of the ethically neutral toss-up belief on the truth of which I am willing to bet either way for the same stakes. But it does not show the independence of the equality of the *value intervals* between F and G and between H and I from that belief. The equality of these intervals depend upon my indifference between (iii") and (iv"). Because indifference is an intensional relation that ranges over the objects of value within its scope, it is not permissible to substitute or logically manipulate the terms of this relation with impunity as Ramsey does and can do with impunity the terms of the equivalence relations set out in (A2). Within the scope of an intensional operator, probabilistically identical toss-up beliefs are not automatically intersubstitutable. For example, I may strictly prefer peaches to pears given an expected 50% chance that a random coin-toss will come up heads. But I may be indifferent between them given an expected 50% chance of rain, because my apprehension about the weather ruins my appetite. So it remains moot whether the value intervals that separate F, G, H and I remain equal relative to some toss-up belief different than *s*.

This raises the question of what it is that axioms (A3) and (A4) stipulate the transitive consistency of. Without an unproblematic derivation of the equality of value intervals FG, GH, and HI that accommodates the intensionality of these values and the relations among them, we lack explicit guidelines for understanding how value options can be equivalent (A3) and how the intervals among them therefore can be equal (A4). We have just seen that an unexplicated transition from indifference-talk to equivalence-talk does not suffice. Then the ordinal ranking of F, G, H and I yields no guidelines for ascertaining what it would mean to speak of the transitivity of "equivalent" value options (A3), nor what it would mean to speak of the transitivity of "equal" value intervals (A4). It appears that the familiar, extensional interpretation of the transitivity relation is the only one available (I examine the Jeffrey-Bolker solution to this problem in Volume II, Chapter III.6.2).

To see why this does not suffice, try replacing the equivalence relation in (A3) with the indifference relation abandoned in the move from (A1) to (A2). Using "≈" to mean "is indifferent to," (A3) becomes

(A3') If $A \approx B$ and $B \approx C$, then $A \approx C$.

(A3′) is not valid even as an empirical generalization, much less as an axiom. For example, (A3′) is violated by my indifference between cherries and apples and between apples and peaches, but strong preference for peaches over cherries. Moreover, there is no way to translate (A3′) into the straightforwardly extensional notation of sentential logic that would preserve the transitive structure of (A3′). Letting P symbolize the sentence, "A is indifferent to B," and Q symbolize "B is indifferent to C," the most we can get out of (A3′) sententially is

(A3″) If P and Q, then R,

which is less than helpful. Since the plausibility of (A4) depends on the suspect transition from (A1) to (A2), it, too, remains suspect. Ramsey's conception of transitivity as spelled out in (A3) and (A4) looks as first glance to be quite innocent, and logically unproblematic. But in both axioms it surreptitiously combines intensional and extensional elements that turn out to be incompatible. Because this conception of transitivity works only by ignoring these underlying incompatibilities, I describe it as *sub-logical*.

It would seem that Ramsey was stuck between a rock and a hard place: either respect intensionality and sacrifice consistency; or ignore intensionality and reap the benefits of sub-logical transitivity. One such benefit was an extensional equivalence relation that obscured not only the intensionality of choice, but thereby the irreducible subjectivity of the chooser – the two persisting obstacles to interpersonal comparisons that, as we saw in Section 1.3, were not circumvented by stipulating that preferences are revealed in behavior. In their absence a cardinal utility scale could be fashioned that might be thought to have extensional application, and so measure the utility-maximization of more than one subject. The price of Ramsey's sub-logical conception of transitivity, however, was the replacement of the primitive notions of preference and utility-maximization he originally set out to axiomatize with an extensional equivalence relation that obtains merely between sentences. Next I show why this Faustian bargain does not circumvent the charge of vacuity.

2.3. Vacuity and Cyclicity

We have just seen that Ramsey's proof presupposes the truth of (U), i.e. that an agent "will act so that what he believes to be the total consequences of his action will be the best possible." We have also seen that the suspect replacement of the indifference relation with the equivalence relation in the move from axiom (A1) to (A2) calls into question whether Ramsey's axioms do, in fact, impose consistency constraints on *bona fide* preference rankings; or whether, instead, they merely impose such constraints on the sub-logical relations among extensional sentences that may or may not assert such

preferences; and whether, therefore, there is in Ramsey's account good reason to venture much beyond (U) to construct a more complex formalization of expected utility theory. But even were these worries about the sub-logical status of these constraints to prove unfounded, there would be independent – though related – reason to move cautiously beyond (U).

An independent implication of (U) is that I always perform that basic action that I believe will effect the best total consequences, i.e. that when I act, I maximize expected utility. This is the vacuous behavioral version of (U) already discussed, according to which the very fact that I perform an action legitimates the inference that I most preferred to perform that action. So from the fact that certain consistency constraints plus Ramsey's motivational assumptions about preference might imply a more complex and formally robust version of (U), it does not follow that lifting these constraints implies the negation of (U). It cannot, because the truth of (U) is assumed at the outset. The concept of preference itself, not that of consistent preference, is what gives meaning to the concept of maximizing a quantity of utility.[23]

---

[23] This point is presupposed in two different methods of evaluating multidimensional alternatives. The *additive model* assigns a scale value to an alternative as a measure of its utility based on the sum of the utility or subjective value of its components. This model satisfies the Ramsey-Savage consistency constraints. The *additive difference model* is based on the difference between the subjective values of two alternatives along some particular dimension. The contribution of this particular difference to the overall evaluation of the alternatives is then determined by a difference function. This model satisfies the Ramsey-Savage consistency constraints only if all difference functions are linear. Although only unidimensional preference rankings will be treated in the following discussion, the significance of these two methods lie in their implication that the utility of alternatives ranked in pairwise comparisons can be established *independently* of the Ramsey-Savage consistency constraints on those rankings themselves. For a discussion, see Amos Tversky, "Intransitivity of Preferences," *Psychological Review 76*, 1 (1969), esp. 41-44.

To my knowledge, the inevitability of intransitive preferences among multidimensional alternatives is first suggested by Ward Edwards, "Probability-Preferences in Gambling," *American Journal of Psychology 66* (1953), 363. In "Intransitivity and the Mere Addition Paradox" (*Philosophy and Public Affairs 16*, 2 (Spring 1987), Larry Temkin ingeniously applies the problem of intransitivity among multidimensional preferences to utilitarianism, in which we must order the number and existence of people and the comparative quality of their lives according to various values such as equality, utility, and the maximin principle (he cites Tversky's paper in a different connection, but nowhere mentions the general intransitivity problem of multidimensional preferences which his discussion illustrates). Ordinarily the problem can be solved *if* all-things-considered judgments about the alternatives can be made that rank them unidimensionally on an ordinal scale. However, Temkin uses the concept of A being all things considered better than B to define "A is preferable to B." This terminology is doubly misleading, for it falsely suggests that multidimensional intransitivities persist in the face of all-things-considered judgments, which they do

Hence the agent can be interpreted as maximizing utility in her actions whether she in fact behaves consistently or not.

Indeed, the concept of expected utility-maximizing preference revealed in behavior would subvert the imposition of consistency constraints on action, even were those constraints not subject to worries about intensionality. For if any action can be interpreted as maximizing the agent's expected utility in virtue of the fact that she performs it, then in particular any preference ranking of alternatives the agent makes at a particular moment can be interpreted as the outcome of consistent pairwise comparisons among all the alternatives available at that moment, and so as reflecting a consistent ordering of those alternatives.[24] Ramsey's result does not redress the vacuity of (U) because it fails to insure the consistency through time of the agent's preferences. As Donald Davidson observes, "The theory merely puts restrictions on a temporal cross-section of an agent's disposition to choose."[25] This means that any behavioral violation of Ramsey's consistency axioms can be understood, in accordance with the principle of charity, as a change in the agent's preferences instead. Hence these restrictions are vacuously inviolable – not because agents never in fact behave inconsistently, but because the Ramsey-Savage concept of a simple ordering[26] by itself is not sufficient to ensure transitivity of preference through time.

---

ordinarily do not; and that all-things-considered judgments express not preference but normative preferability, which it ordinarily does not. Temkin's discussion is valuable because of its breadth, detail, and focus; but these terminological idiosyncrasies lend it a greater air of paradox than seems warranted. I suggest a way of avoiding intransitivities caused by preferences among multi-dimensional alternatives in Volume II, Chapter III.9.

[24] I use the term "preference ranking" to refer to the result of a pairwise comparison between two given alternatives, and "ordering" to refer to the resultant relations of priority that obtain among all such alternatives consecutively ranked. One way of putting my point would be to say that the burden of interpreting the concept of maximizing utility is carried by the concept of a preference ranking, not by that of a simple ordering. This is why the impossibility of linearly representing intransitive preferences does not exclude their cyclical ordering. For an example, see the discussion of Cleopatra in Section 3.2, below.

[25] Donald Davidson, "Psychology as Philosophy," in *Essays on Actions and Events* (Oxford: Clarendon Press, 1980), 235. Sen makes essentially the same point in "Rational Fools: A Critique of the Behavioral Foundations of Economic Theory," *Philosophy and Public Affairs 6*, 4 (1977), 317-44; see esp. 325.

[26] As refined by Leonard Savage in *The Foundations of Statistics* (New York: Dover Publications, Inc., 1971), 17-21; also see R. D. Luce and Howard Raiffa, *Games and Decisions* (New York: John Wiley and Sons, Inc., 1957), 25-28. Tjalling Koopmans (in "Allocation of Resources and the Price System," in *Three Essays on the State of Economic Science* (New York: McGraw-Hill, 1957) dismisses the time-dependence problem

To sharpen the underlying problem here, I consider examples using the strict preference relation, rather than the indifference relation on which Ramsey relied, to construct a scale of equal value intervals. Following Ramsey, let ">" be the relation, "is strictly preferred to." Then a *transitive ordering* of the kind adopted by Savage and others of alternatives F, G, H, ... must conform to the following *rule of transitivity*:

(T) If F>G and G>H, then F>H.

(T) is where the intensionality problem and the vacuity problem begin to dovetail. We have already seen, using the indifference relation as an example, that it is not possible to cast a transitivity relation among intensional preferences in a strictly sentential form. This is even truer for examples using the relation of strict preference. Hence (T) is regarded as what Savage calls a "logic-like" criterion of consistency among alternatives in decision-making.[27] It is not strictly a criterion of logical consistency because it is neither among nor implied by the laws of logic. Nor can we conclusively identify the relation between (T)'s antecedent and consequent as one either of logical entailment or material implication.

If (T) *were* an axiom of logic, then it would assert *something like* a conceptual truth about what it means to prefer F to G and G to H. In this case, to violate (T) would be to have no genuine preferences among F, G, and H at all. If, on the other hand, the relation between (T)'s antecedent and consequent were that of material implication, then it would assert a merely truth-functional relation between preferring F to G and G to H on the one hand, and F to H on the other. In this case, to violate (T) would be simply to furnish an instance that merely falsifies (T), in which one's genuine preferences happen to be intransitive. Utility theorists vacillate between treating (T) as though it involved something like material implication, and as though it involved something like logical entailment.[28] But presumably we would think (T) – or

---

discussed below by stipulating that on one interpretation of his proposed model of competitive equilibrium, "a choice by a consumer is in fact a plan for future consumption extending over all periods considered. His preference ordering is thought of as an ordering of all such plans" (61). But this will not suffice if any such meta-ordering realistically may change from moment to moment as the consumer herself advances in time. By contrast, Donald Davidson, J. C. C. McKinsey, and Patrick Suppes suggest that we preserve transitivity by reinterpreting changes in preference over time as changes in the alternatives ranked. This makes (U) *unnecessarily* vacuous, since they have already denied any interest in how people may in fact order their preferences. See their "Outlines of a Formal Theory of Value, I," *Philosophy of Science 22* (1955), esp. 144-5.

[27] *Ibid*., Savage, *The Foundations of Statistics*, 19.

[28] For example, I. M. D. Little (in "A Reformulation," *op. cit.* Note 21) and Davidson, McKinsey, and Suppes (*op. cit*. Note 26) treat (T) as involving entailment, whereas Ward

something like it – rational, even if no one's behavior ever conformed to it. So whatever (T) is, it is not an empirical generalization describing actual choice behavior. This should be kept in mind in what follows.

(T) has sufficient structural and intuitive similarity to certain laws of logic – transitivity of implication, for example – that we tend prereflexively to agree with Ramsey in thinking "inconsistent" (in some sense) an agent whose choice behavior violate (T) in a *cyclical ordering* of F, G, and H:

(C) F>G and G>H and H>F

(C) seems in some sense inconsistent because we reflexively think, first, that F>G and G>H "imply" F>H; and second, that F>H and H>F are mutually contradictory. But although implication and mutual contradiction are well-defined within the apparatus of classical logic, it is not possible to give them a similarly rigorous meaning here. Thus because H>F does not *logically* imply not-F>H, F>G, G>H, F>H, and H>F all may be true together. Hence (T) and (C) may be, too. Because Ramsey's value axioms secure (at best) the consistency of value measurement units and value intervals among ranked alternatives and not the mutual consistency of the pairwise comparisons among alternatives on which such a ranking is based, they do not exclude (C); nor, therefore, the sub-logical "inconsistency" that (T) and (C) represent.

Whereas (T) is not an empirical generalization over actual behavior, (C) appears to be. As Sen has shown, the theory of revealed preference preserves the distinction between preference and "selection"[29] behavior. This means that intransitive selection behavior described by (C) need not violate the transitive preference ordering described by (T), any more than an agent who sequentially asserts inconsistent beliefs thereby commits a logical impossibility. Just as we save the assumption of rationality in an agent whose speech behavior calls it into question by attributing to her rational beliefs that conform to the laws of logic, we similarly may save the assumption of rational preferences in an agent whose selection behavior is intransitive by attributing to her preferences that conform to (T).

Edwards (in "Probability-Preferences in Gambling" (*op. cit.* Note 23) treats it as involving, at best, material implication.

[29] Thus I adopt Ullmann-Margalit and Morgenbesser's nomenclature (see Edna Ullmann-Margalit and Sidney Morgenbesser, "Picking and Choosing," *Social Research 44*, 4 (Winter 1977), 757-785). But I use "selecting" as the generic term in order not to prejudge the question whether every selection is, in their terminology, a choice (hence reflects a preference), rather than in order to raise the question whether picking is a real possibility. I am convinced by Ullmann-Margalit and Morgenbesser that it is, but my account addresses the canonical preference (-or-indifference) relation in order to preserve the completeness condition on orderings.

So the quasilogical status of (T) can be retained, and the empirical findings respected, by invoking the principle of charity.[30] Rather than charge the agent whose selection behavior is described by (C) with inconsistency, we may instead simply revise our hypothesis about her present preference rankings. Thus suppose Cyril's behavior produces a cyclical ordering over three temporally sequential trials, in each of which he must make pairwise comparisons among F, G, and H, as follows:

(C_t) t_1: F>G
t_2: G>H
t_3: H>F

Call these temporally sequential trials *time-dependent*. That Cyril prefers H to F at t_3 permits us to infer that he has changed his mind about his earlier rankings at t_1 and t_2, i.e. that at t_3, he rather prefers H to G and G to F. The vacuity of the Ramsey-Savage concept of a simple ordering arises from the fact that the transitivity of an agent's preferences always can be preserved by making this inference. Because all selection behavior always permits it, no such behavior, not even that described by (C), can be shown to violate (T).[31]

---

[30] For extended discussion of this principle, see W. V. O. Quine, *Word and Object* (Cambridge, Mass.: M. I. T. Press, 1960), 59, 69; *Ontological Relativity and Other Essays* (New York, N. Y. Columbia University Press, 1969), 46; Donald Davidson, "On the Very Idea of a Conceptual Scheme," APA Presidential Address, *Proceedings and Addresses of the American Philosophical Association 47* (1974). Also see Amos Tversky and Daniel Kahneman, "Judgment Under Uncertainty: Heuristics and Biases," *Science 185* (1974), 1124-31; ""The Framing of Decisions and the Psychology of Choice," *Science 211* (1981), 453-458; James G. March, "Bounded Rationality, Ambiguity, and the Engineering of Choice," *Bell Journal of Economics 9* (1978), 587-608; Elliot Sober, "Psychologism," *Journal for the Theory of Social Behavior 8* (1978), 165-191; L. Jonathan Cohen, "On the Psychology of Prediction: Whose is the Fallacy?" *Cognition 7* (1979), 385-407; "Can Human Irrationality be Experimentally Demonstrated?" *Behavioral and Brain Sciences 4* (1981), 317-331; Steven P. Stitch and Richard E. Nisbett, "Justification and the Psychology of Human Reasoning," *Philosophy of Science 47* (l980), 188-202; Richard E. Nisbett and Lee Ross, *Human Inference: Strategies and Shortcomings of Social Judgment* (Englewood Cliffs, N. J. Prentice-Hall, 1980); Paul Thagard and Richard E. Nisbett, "Rationality and Charity," unpublished paper, 1981; Steven P. Stich, "Could Man be an Irrational Animal?" *Synthese 64*, 1 (1985).

[31] In "The Theory of Decision-Making," (*Psychological Bulletin 51*, 4 (1954), Ward Edwards acknowledges and criticizes this argument on the grounds that "unless the assumption of constancy of tastes over the period of experimentation is made, no experiments on choice can ever be meaningful, and the whole theory of choice becomes empty. So this quibble can be rejected at once" (405). !!? These concluding sentences do not follow as obviously as Edwards thinks they do. My argument in this chapter is that, on the

Since the appeal to the principle of charity is intended to interpret a temporally sequential series of pairwise comparisons of alternatives, and since this appeal is required in order to preserve the transitivity of those comparisons in the face of (C), the vacuity of the concept of a simple ordering arises, in part, from its implicit time-dependence.[32] Then even when (C_t) obtains, (U) is satisfied; i.e. (U) is vacuous.

Seemingly intransitive behavior explained by systematic changes in taste resulting from learning or sequential effects over repeated trials is one obvious application of the principle of charity. But (T) may be vacuously preserved by appeal to this principle even in cases otherwise explained as a momentary lapse or glitch in the evaluative process. In these cases, too, no genuinely intransitive behavior can be identified.[33] Thus the probabilistic

---

contrary, the utility-maximizing theory of choice may meaningfully survive if its explanatory scope is reduced. If not, I think it is not "this quibble" that should be "rejected at once." Davidson, McKinsey, and Suppes (*op. cit.* Note 26) conceive the problem as consequent on a revealed preference interpretation of choice. They think that if preference is taken to be equivalent to selection behavior, then each time-dependent selection may spring from a "momentarily rational preference ranking," and so we cannot prove *what* a person's preference ranking is over more than two alternatives. Their solution is to interpret particular selections as *evidence* for preference interpreted as a disposition to select. A cyclical ranking is then "evidence, so far as it goes, that [a person's] preference ranking is not rational; but we would reconsider this verdict if we learned he had changed his mind about the relative ranking of *a* and *c* after his first two choices" (147). But to change one's mind after each selection just is to have a series of "momentarily rational preference rankings." So the dispositional interpretation of preference is of no help here; I address some further arguments for this interpretation in Note 34, below. Also see Donald Davidson, Sidney Siegel, and Patrick Suppes, "Some Experiments and Related Theory on the Measurement of Utility and Subjective Probability," Applied Mathematics and Statistics Laboratory, *Technical Report 1*, Stanford University, Stanford, Cal., August 15, 1955.

[32] Notice that it will not solve the problem to devise a way for Cyril to select a simultaneous linear ordering of F, G, and H – say, by requiring him to choose among pressing buttons for an F-G-H, an H-G-F, and an F-G-H-F scale. If he chooses the last, the time-dependent appeal to the principle of charity can be made for scanning the scale just as for pairwise-ranking alternatives. The implications for the use of the principle of charity of "changing one's mind" under such circumstances are discussed further in Section 3.1, below. In any case, a simultaneous linear ordering is a plausible substitute for pairwise comparisons only where the number of alternatives to be ordered is sufficiently restricted to those which are simultaneously comparable in practice. Even three or four such alternatives, simultaneously presented, is a stretch.

[33] Again I mean to be referring only to unidimensional criteria for ranking alternatives. In "Intransitivity of Preferences" (*op. cit.* Note 23), Tversky demonstrates that genuinely intransitive behavior can be identified when multidimensional criteria are used. He acknowledges, however, that in the absence of replication, "one can always attribute intransitivities to a change in taste that took place between choices" (45). For this reason

reformulation of (T) designed to incorporate such intransitivities, *weak stochastic transitivity*, or WST, is susceptible to similar complaints. Let the probability p(F,G) of choosing F over G in a choice between them and the probability p(G,F) of similarly choosing G over F equal 1. Then WST defines the preference for F over G thus:

(WSP) F>G if and only if $p(F,G) \geq 1/2$,

i.e. F is defined as preferred to G if and only if it is chosen over G more than half the time. Then the rule of weak stochastic transitivity can be restated as follows:

(WST) If $p(F,G) \geq 1/2$ and $p(G,H) \geq 1/2$, then $p(F,H) \geq ½$

But by elevating preference to a second-order probabilistic function over a range of transitive and intransitive choices, WST replaces revealed preference – the original target of the argument for (U)'s vacuity – with a new target, namely WST's unanalyzed notion of choice, which is susceptible to the same critique.[34]

---

he introduces variables into the experimental design "to minimize the memory of earlier choices in order to allow independent replications within one session" (34). Presumably, the point of minimizing memory from one trial to the next is to minimize the possibility that the agent's choice is motivated by a desire to maintain consistency, rather than to maximize utility. This rationale begs my question, of *whether there can be* a coherent concept of utility-maximization that does not presuppose conformity to the requirements of logical consistency. (An example of such conformity would be L. L. Thurstone's "naive" subject with an "even disposition", whom he instructed to assume a "uniform motivational attitude" (personal communication cited in Ward Edwards, "The Theory of Decision-Making," *op. cit.* Note 31. See L. L. Thurstone, "The Indifference Function," *Journal of Social Psychology 2* (1931), 139-167).) It also begs the question of what an authentic choice is. Tversky seems to assume it is a matter of reflexive, unreflective behavior, rather than an expression of conscious and consistent deliberation about all the relevant considerations. But it is unclear why independent replications based on minimal memory of previous rankings should be treated as more authentic expressions of preference than those influenced by memory or by a hypothesized change of mind. I air some further reservations about this device in Section 3, below.

[34] Can this problem be avoided by defining preference rankings dispositionally rather than time-dependently? Suppose we stipulate that the agent has intransitive preferences if at any t_n, she is disposed *simultaneously* to select F over G, G over H, and H over F. On the basis of empirical evidence, a counterfactual account could then be given of how she would select time-dependently if confronted with pairwise comparisons among F, G, and H. The empirical evidence for this might consist in the following sort of statistical study:

The *money pump* may seem to refute (U)'s vacuity. A money pump is an agent who pays out – hence loses money – for the privilege of making cyclical preferences in the manner of (C_t). Thus suppose Hazel at t_1 purchases F, a cell phone/ alarm clock/ radio/ video camera/ gameboy/ pager with maximal roaming flexibility for $200.00. At t_2 she pays a fee of $10.00 to exchange F for G, a cell phone/ alarm clock/radio/video camera/pager with no gameboy (because she thinks it's silly) but maximal roaming flexibility for $150.00. At t_3 Hazel pays another fee of $10.00 to exchange G for H, a basic cell phone with no accoutrements and sufficient roaming flexibility for her needs for $50.00, because the video camera freezes up when the pager or alarm clock is

---

> We explain to each of 1,000 subjects that, given F and G, if they reach out a hand for F, G will be removed. We describe these subjects as "selecting F" if they reach out for F under these circumstances. We then give to each subject the same instructions regarding G and H, and F and H: Whichever the person reaches out her hand for is that which we interpret her has having "selected." We find that, over a series of such trials, randomized over individual subjects, this population selects F over G, G over H, and H over F.

On this basis, when Hilda later selects F over G at t_1, G over H at t_2, and H over F at t_3, we can then infer that at t_3 she is disposed to select F over G, G over H, and H over F. We conclude that Hilda has genuinely intransitive preferences. She has not merely changed her mind about her earlier ones.

However, the statistical study legitimates this conclusion only if we have evidence that at least most subjects in the test population have not changed their minds about their earlier preferences. At most, their behavior encourages the generalization that when people select F over G and G over H, they often select H over F. But this generalization is neutral between their changing their minds and having intransitive preferences. It is hard to imagine what empirical data would justify the above conclusion, without arbitrarily ruling out even the possibility that Hilda had changed her mind.

Second, the notion of a disposition that is intended to explain Hilda's cyclical ordering itself encounters the same problems as the theory of revealed preference. If "disposition" means the structural propensity (causal, teleological, statistical) to behave as she did in fact behave, the disposition at t_3 to select G over G, G over H, and H over F is consistent with the hypothesis that she selected H over F at t_3 because she changed her mind about her earlier preference rankings; i.e. that Hilda now at t_3 prefers G over F and H over G – regardless of how she is structurally disposed to behave. Either it is vacuously true that people always prefer what they are disposed structurally to select – in which case there is no significant difference between preference and selection behavior after all; or else such behavior must be supposed to "reveal" "preference" in some more full-blooded psychological sense that is independent of behavior. And then, as Sen's argument suggests, it is an open question whether one's behavior reveals that state or not. So the vacuity of (T) is not mitigated by invoking dispositions as an alternative to time-dependent behavior. I owe this objection to Allan Gibbard.

beeping. But at t_4 Hazel pays a third fee of $10.00 to exchange H for F because F maximizes her options. At t_5 she pays a fourth fee of $10.00 to exchange F for G because she thinks the gameboy is silly. And so on. By now Hazel has paid $40.00 for the privilege of ending up with her original choice and embarking once again on the cycle of preferences, and there is no end – aside from Hazel's bankruptcy – in sight. The argument would be that since the money pump depletes her own resources with no noticeable return, she clearly fails to maximize utility, and so demonstrates the irrationality of a cyclical ranking.

However, both premises of this argument are false. Changing her mind repeatedly among the available options really is a privilege – and not only that but a luxury – for which Hazel is quite rightly willing to pay. When she is feeling expansive and adventurous, she prefers F; and both G and H would be wrong for her in that mood. Similarly, when she is feeling like a Serious Person, she prefers G; and both F and H would be wrong. And when she is feeling austere and disciplined, she prefers H; F or G wouldn't do at all. Hazel is paying for the privilege and the luxury of expressing her shifting moods in her choice of cell phone. What's so irrational about that? – Nothing, according to (U). Although the concept of utility has, as we have seen, many possible interpretations, it cannot be meaningfully distinguished from notions of desire-satisfaction, pleasure, or preference, however loosely these are interpreted. But we also have seen in Chapter II.2.3 that the standpoint of desire affords no emotionally detached perspective, external to the agent's system of desires overall, from which the agent might evaluate the worth of this system itself – much less the cyclical preferences it may contain, or the quality of those preferences. (U) tells us that if we wish to express our shifting moods through our shifting choices of cell phone, and have the means to satisfy this wish, there is no reason not to do so.

3. Transitivity

The time-dependence of a simple ordering generates a dilemma about when to invoke the principle of charity to protect (T) against seemingly intransitive preferences. Below I argue that protecting (T) using the principle of charity must defer to a more important rationality assumption, i.e. psychological consistency, that may be incompatible with (T) under certain circumstances. In the next subsection I argue that psychological consistency presupposes logical consistency; and therefore that under those same circumstances, (T) and logical consistency are incompatible as well.

3.1. Minimal Psychological Consistency

Suppose that F, G, and H represent, not relatively humble, conceptually pedestrian alternatives like apples, oranges, and pears, but rather more inclusive, personally momentous, but often mutually incompatible alternative

meta-ends like being honest, being entertaining, and being tactful, respectively; and that Amos gives these three a cyclical ordering as described by (C). To suppose that Amos has merely changed his mind at t_3 is to suggest that he can vacillate with ease among the values of honesty, entertainment, and tact, just as he might among apples, oranges, and pears; i.e. that adopting and discarding these behavior attitudes, and all that each implies for the more specific choices of behavior he must make from moment to moment, is no more problematic than adopting and discarding a jacket when the weather changes. An agent capable of effortlessly altering his primary characterological priorities from one moment to the next, or whenever new alternatives and circumstances present themselves, either purchases transitivity at the price of psychological consistency, or at least gives new meaning to the concept of spineless compromise.

But this is to conceive only the more realistic case. Clyde, who reorders literally all of his priorities from one choice occasion to the next, presents the more radical challenge to the principle of charity. By hypothesis, Clyde is able to sustain an enduring sense of psychological continuity at the same time that all his preferences undergo revision from moment to moment. But in the absence of at least some enduring priorities, it is not easy to see in what this sense of psychological continuity could consist. Under these circumstances, Clyde would be little more than an enduring physical entity, constantly bombarded by new possibilities, constantly changing his mind about what to do and what is important, with no psychological consistency from one moment to the next. We can assume, for the sake of argument, that he has memories of previous ranking occasions. Indeed he must, in order to be capable of remembering, from moment to moment, what the enterprise of ranking alternatives requires him to do. But by hypothesis, he is unmotivated to recall these previous orderings as in any way important or interesting, for there are no ongoing goals, values, or plans of action of overriding importance relative to which they can be assessed and ranked.

But the problem is even worse than this. So far I have described Clyde's dilemma as one concerning choice among alternatives that can be consistently ordered. However, Clyde himself is incapable even of picking among alternatives all of which are acceptable. I quote at length Edward McClennen's description of Clyde's dilemma:

> How is he to pick? Suppose that he decides to settle it by the flip of a coin: if heads, he will pick *x*, and if tails, he will pick *y*. Let him now perform the experiment and observe its outcome. Whatever the outcome [heads or tails], why now should that outcome settle anything as to which one to pick? The decision to settle the matter by the toss of a coin is history. … Moreover, it is still the case that from a [utility-maximizing] perspective he has no basis for deciding which one to pick. Perhaps he should flip the coin again! Alternatively, suppose that [Clyde] simply

> finds himself reaching for *x* rather than *y* and then, in the middle of the reach, the thought crosses his mind to reconsider – not to reconsider the evaluation that led to the determination that both *x* and *y* are fully acceptable, but to reconsider the settled picking of *x* instead of *y* that the reach toward *x* implies. From a [utility-maximizing] perspective, there is still no basis for the picking of *x* rather than *y*. Both are still open to him. Whatever impulse it was that resulted in the agent's hand reaching toward *x*, that impulse, given the intervening reflection, is now history.[35]

In such cases, it is difficult to comprehend the sense in which Clyde could be understood even as picking, much less as choosing anything at all. Hence Clyde would be an agent only in a truncated sense, for he would be unable even to initiate, let alone carry out any sustained plan of action.[36]

Earlier it was argued that we could preserve under all circumstances the hypothesis that an agent's preferences are transitive, by appealing to the principle of charity to ground the assumption that seemingly intransitive selection behavior in truth represented the agent's changes of mind about her earlier preference rankings. This seemed in keeping with the canonical practice of accepting that explanation of an agent's behavior that preserves the assumption of her rationality. But Clyde's is a cautionary tale that warns us against protecting the rule at the expense of the rationality of the agent. For it appears that in this extreme case the principle of charity preserves (T) only by undermining what we can now see is an even more important rationality assumption of *minimal psychological consistency*, i.e. that a rational agent retain some long-term priorities that are not subject to constant revision in ranking status. It is more important that Amos and Clyde retain some such long-term priorities, than that their moment-to-moment selections among unidimensional alternatives be transitive; and these two rationality assumptions may conflict: Clyde's selection behavior can be (repeatedly re)interpreted as transitive only if he is assumed to lack psychological consistency. (T) as it stands does not protect the assumption of psychological consistency that is necessary for agency, and so we doggedly protect (T) at its peril. Since the standard reading of (T) may force an unacceptable choice between transitivity and psychological consistency, we need some way of

---

[35] Edward McClennen, *Rationality and Dynamic Choice: Foundational Explorations* (New York: Cambridge University Press, 1990), 208. I discuss this passage at greater length in Volume II, Chapter IV.6.

[36] The psychological consistency provided by ongoing goals, values, and plans that forms the focus of this and the next section is usually presupposed in discussions of personal identity. See, for example, the essays by Locke, Grice, Perry, Williams, and Parfit in John Perry, Ed. *Personal Identity* (Los Angeles: University of California, 1975). Also see Joel Feinberg's analysis of Durkheimian *anomie* in "The Idea of a Free Man," in *Rights, Justice, and the Bounds of Liberty* (Princeton: Princeton University Press, 1980).

incorporating this requirement into a more sophisticated interpretation of (T) that circumvents it.

As stated, the requirement of psychological consistency is an extremely weak one to impose on a transitive preference ordering. For example, Doris, who at t_1 prefers her galoshes to her sneakers, at t_2 her sneakers to her high heels, and at t_3 her high heels to her galoshes does not violate it – provided that she also prefers both at t_1 and at t_2 that she try to make the world a better place, at t_2 and t_3 that her loved ones prosper, and so on. That is, it takes only one ranking priority that is unrevisable from t_1 to t_2 in order to satisfy the minimal requirement of psychological consistency as stated here. Nor does this requirement exclude Cecil, whose only stable, nonrevisable ranking priority is his sneakers. That an agent's nonrevisable ranking priorities must be psychologically fundamental rather than frivolous or peripheral in order to sustain psychological consistency in some more full-blooded sense is a thesis I shall not defend, though I believe it is true. The minimal requirement of psychological consistency that I now examine more closely is merely a skeletal adumbration of a more complex and realistic one that I do not think it necessary to develop here.

Satisfying the requirement of minimal psychological consistency is a necessary condition for intentionally ranking alternatives. Suppose Winifred must rank three alternatives F, G, and H in a series of three pairwise comparisons, and that her other preferences are, as a matter of psychological fact, transitive and temporally stable. In order for Winifred's selection behavior in this instance to count as an instance of intentionally ranking F, G, and H, she must both

> (a) have the concept of something's ranking superiority, i.e. the concept of its being preferred to other available alternatives; and also
>
> (b) remember on each trial the relation of the pair she is ranking to the third alternative she is not.

That is, when ranking F and G on the first trial, she remembers that H has yet to be ranked; when ranking G and H on the second trial she remembers that F has already been ranked; and similarly with G when ranking F and H on the third trial.[37]

---

[37] (b) has been regarded as much more controversial than it should be. Tversky's version of the money pump seems to depend on ignoring it (Amos Tversky, "Intransitivity of Preferences," *op. cit.* Note 23, 45); and Thomas Schwartz ("Rationality and the Myth of the Maximum," *Nous 6* (1972), 97-117) is surely right in observing that "once you realize you are engaged in a drawn-out process of choosing among three options, even though only two are available at any one instant, you should spend your money as if you are choosing among all three options – paying for a given option only if it is optimal among

Now suppose a series of contiguous moments in time during which Winifred is to rank F, G, and H. At t_1 she is presented with alternative F and G and selects F. At t_2 she is presented with G and H and selects G. At t_3 she is presented with alternative F and H. Now assume she already has the concept of something's ranking superiority, i.e. has ranked alternatives on previous occasions and learned to conceptualize her behavior in the usual way (a), and remembers the rankings she has already given to F relative to G and G relative to H (b). Then Winifred can conclude that F is superior in ranking not only to G, but to H as well. Her memory of her earlier rankings in effect establishes her third. So selection behavior that satisfies conditions (a) and (b) for intentionally ranking alternatives thereby satisfies (T). Winifred has at least one relatively long-term priority throughout each of three ranking trials. This establishes her as minimally psychologically consistent and her preference ranking as therefore transitive. Although we have already seen that having transitive preferences is compatible with psychological inconsistency, an agent who is minimally psychological consistent will have at least one transitive preference ranking.

Contrast Winifred with Rex, who has and at t_3 applies the concept of ranking superiority (a) to H, because he has thoroughly forgotten the relation of F and H to G established by his two previous rankings. That is, Rex's selection behavior satisfies (a) and violates (b). Then he can, *contra hypothesi*, draw no conclusions as to the ranking superiority of any one of the three alternatives to any of the others. That is, if

(1) t_1: F>G
(2) t_2: G>H
(3) t_3: H>F,

then by transitivity,

(4) G>F (on (2) and (3))
(5) H>G (on (3) and (1))
(6) F>H (on (1) and (2)).

So Rex must conclude that everything is preferred to everything else, hence that none of the three alternatives is superior in ranking to any of the others. So, in particular, it is not superior in ranking to F. So his application of the concept of something's ranking superiority to H at t_3 has involved him in a

---

the three and you cannot make an optimal choice without paying as much or more" (109). I have more to say about this below.

straightforwardly *logical* inconsistency: H both is and is not preferred to F.[38] Just because we cannot symbolize a logical contradiction within the constraints of standard decision-theoretic notation does not mean we do not know it when we see it, nor that we cannot describe it in natural language. We have just done so. So although Rex has thereby produced a cyclical ordering, he nevertheless cannot be said to have ranked F, G, and H at all.[39] Rex's case shows that, as Kant might have put it, concepts without memory come up empty. Without a recollection (or "synthesis") of previous ranking occasions to supply continuity with this one, the concept of a thing's ranking superiority is applied so indiscriminately that it has no proper application at all.

But now contrast Winifred with Wallace, whose selection behavior violates (a) but satisfies (b). Wallace remembers on each trial the relation of the pair of alternatives he is ranking to the third he is not, but lacks the concept of ranking superiority to apply to his behavior. If he never forms this concept, he cannot be described as preferring any one alternative to any other. But could Wallace develop and consistently apply that concept to the ordering he produces? – Yes, assuming he does not thoroughly change his mind about his previous orderings. That is, he can form the concept of ranking superiority only if he is minimally psychologically consistent. Minimal psychological consistency may provide a kind of foundational structural support that determines the transitive consistency of future pairwise comparisons with past ones, and so enables the concept of ranking superiority to develop. In the absence of this support, it is hard to imagine how it might.

Take Wallace's first pairwise comparison, the choice of F over G at t_1. This alone would not enable him to form the concept of F's ranking superiority. For this would be to treat F as an instance of something's ranking superiority. But

---

[38] Essentially this is Davidson, McKinsey and Suppes' defense of transitivity (*op. cit*. Note 26, 145-6), although they do not distinguish unidimensional from multidimensional orderings. They recognize that the irrationality of a cyclical ordering consists not simply in a violation of transitivity, but of *logical* consistency in the application of the concept of rational choice. That concept plays the same role in their argument that the concept of ranking superiority plays in mine in this volume; in Volume II, Chapter III, I demand greater formal rigor from the notion of logical consistency. Here Schwartz (*ibid*.) seems patently mistaken in supposing that under these circumstances, "it is a matter of indifference" which alternative is chosen – unless he means it is a matter of indifference to an agent who is unable to comprehend what is involved in ranking alternatives at all.

[39] Of course if Winifred forgets the relation of the pair she is ranking to the third alternative she is not, she is free to apply the concept of ranking superiority locally, to either of the two alternatives presented on each of the three trials sequentially, without regard to what has occurred on either of the others. But it will still be true in fact that, at t_3, when she produces a cyclical ordering of F, G, and H, we will have applied that concept inconsistently, if we continue to insist on describing Winifred's behavior as an instance of intentionally "ranking F, G, and H."

as yet Wallace has no such concept for F to be an instance of, and he can't form an inductive generalization on the basis of a single case. Next proceed to the second pairwise comparison, in which Wallace chooses G over H at t_2. It might seem that we now have the second case on which the inductive generalization to the concept of something's ranking superiority can be made. But we do not. For the concept, "something's ranking superiority," is identical to the concept, "some (one) thing's ranking superiority;" and there is no one thing at both t_1 and t_2 from which that concept can be inductively derived. Again: of course both F at t_1 and G at t_2 are instances of this concept. But the question remains of how Wallace might ever come to form it in the first place. Call this *Wallace's learning problem* (without, however, intending to suggest that the problem is Wallace's). He comes to form it by remembering at t_2, while choosing G over H, the alternative he chose over G at t_1, namely F. The "some one thing" that has ranking superiority both at t_1 and at t_2, therefore, is F. The concept of a thing's ranking superiority is an inductive inference from Wallace's memory of his first two pairwise comparisons. His ranking of F over H at t_3 then follows by implication.

This conclusion assumes that, while making the third pairwise comparison between F and H at t_3, Wallace has not changed his mind about his ranking of F and G at t_1 and his ranking of G and H at t_2. This is where minimal psychological consistency becomes important. For if he now at t_3 were to prefer G to F and H to G, the effect would be the same as if he had not changed his mind but instead violated the consistency of his earlier rankings by choosing H over F: He would have produced a cyclical ordering, and there would no longer be some one thing ranked as superior in at least two cases over which to generalize inductively. But we have already seen that, in this event, the concept of a thing's ranking superiority would be otiose even if he had acquired it. Kantians will recognize this argument as a streamlined version of Kant's analysis of transcendental synthesis at A 99 – A 102.[40]

So far the argument has been that intentionally ranking a set of given alternatives is possible only if two necessary conditions are satisfied:

> (a) The agent must be able to form and apply consistently over time the concept of a thing's ranking superiority; and
>
> (b) she must remember the relation of the two alternatives she is presently ranking to the third she is not.

[40] Immanuel Kant, *The Critique of Pure Reason*, trans. Paul Guyer and Allen W. Wood (New York, N.Y.: Cambridge University Press, 1998)

This is what it means to have a genuine preference. A genuine preference can be expressed in a series of pairwise comparisons that satisfy (T). But a genuine preference also expresses an agent's minimally psychologically consistency.

Now consider briefly what is involved in the requirement of psychological consistency, i.e. that an agent have at least some long-term priorities that are not subject to constant revision in ranking status from one pairwise comparison to the next. First, she must have the concept of a thing's priority in her system of ends, and be able to reapply it to those alternatives she ranks most highly from moment to moment. This just is the concept of a thing's ranking superiority (a) just discussed. Second, she must remember from moment to moment that her long-term priorities are not subject to constant revision. This means that she must remember the relation of these priorities to any two alternatives she is presently ranking. They must constitute a third alternative relative to any pairwise comparison she is presently making, such that her long-term priorities actually are prior relative to other alternatives available, i.e. she assigns them most preferred status relative to these other alternatives. This is equivalent to (b), above. (a) and (b) together imply that her ordering of the given alternatives is transitive. I discuss the significance of (a) and (b) for the more comprehensive, Kantian model of rationality in which utility-maximization must be embedded in Volume II, Chapter III.

Of course there are other conditions that must be satisfied in order that an agent be psychologically consistent. For example, she must be individuated, capable of conceptual discrimination, and so forth. But these conditions must be satisfied in order that she be able intentionally to rank alternatives as well. So an agent can intentionally and consistently rank a given set of alternatives – i.e. express a genuine preference – only if she is psychologically consistent at least in the minimal sense described.

In the following subsection I propose, following Kant, that psychological consistency presupposes logical consistency, in which case the concept of a genuine preference does, too. I also suggest briefly some desiderata that a revised version of (T) would need to satisfy. This completes the more general, critical part of the argument that (T) must be subordinated to the requirements of logical consistency in order to avoid the unappetizing alternatives of vacuity and inconsistency. I offer a substantive and detailed picture of what (T) looks like when thus subordinated in Volume II, Chapter III of this project.

3.2. Logical Consistency

Winifred's and Wallace's cases demonstrate that having and consistently applying the concept of a genuine preference is a necessary condition of observing the rule of transitivity; and so the necessity that an agent possess the concept of a genuine preference when intentionally ranking alternatives.

This is a condition about which conventional decision-theoretic interpretations of (T) have nothing to say.

(T)'s silence on the concept of a genuine preference results not only from conflating genuine preference with selection behavior, as Sen has shown. It therefore results from failing to incorporate the fact that, like any human behavior, the selection behavior of making pairwise comparisons is physically discrete and time-dependent, whereas the logic and concepts that inform the preferences thereby expressed are not. The process of making transitive pairwise comparisons can be represented as selection behavior of the following sort, where "s" means "is selected over":

(T_t) t_1: FsG
t_2: GsH
t_3: FsH

Now ordinarily, when Horace is asked to rank F, G, and H, we assume (b) above, i.e. that he remembers at t_2, while ranking G and H, his previous ranking of F and G at t_1. So if he has selected F over G at t_1 and G over H at t_2, we do not therefore suppose, as we did provisionally for Wallace, that Horace has changed his mind about G from t_1 to t_2, first ranking it as inferior and then as superior. We do not conclude this because we understand at t_2 that whereas earlier Horace was ranking G relative to F, he is now ranking it relative to H; and these two rankings are obviously consistent. Ordinarily, then, we also assume condition (a) above, i.e. that there is a most selected alternative common to both trials, namely F. These commonsense assumptions ensure the interpretation of Horace's preferences as psychologically consistent.

Canonical preference language gives no indication of this psychological consistency, and fails to incorporate the conditions that ensure it. We have already seen that this is what makes (T) vacuous. Instead, it collapses preferences that satisfy (a) and (b) into a time-dependent series of discrete physical selection behaviors. Thus literally interpreted, the first two pairwise comparisons *are* psychologically unconnected and temporally sequential, such that G is ranked lowest in the first trial and highest in the second (recall that this absence of a most highly ranked alternative common to t_1 and t_2 accounted for Wallace's learning problem). But of course this literal interpretation is not the one we make, for it misrepresents the rank we assume Horace has in fact assigned to G. We do not think that he prefers G most at one moment and least the next. This *would* be logically inconsistent, in the familiar, time-dependent sense in which we say of someone with unstable or shifting opinions that they are "constantly contradicting themselves". Describe this as *intertemporal logical inconsistency*. If Horace were simply to

prefer an alternative most at one moment and least the next, his preference would be intertemporally logically inconsistent. In that case he would have no genuine preference at all.

In interpreting the canonical language of preference, we instead make the correct but extranotational assumption that in selecting F over G at t_1 and G over H at t_2, Horace is applying a time-independent, logically consistent *rule*, namely the concept of a genuine preference, in the selection behavior he enacts. (T)'s horseshoe should be understood as expressing the *conceptual implication* that by ranking F over G and G over H, one *thereby* ranks F over H, and so expresses a genuine preference. On this reading, (T) implicitly expresses a conceptual truth. An agent like Horace who has a genuine preference for F over H will be constrained by the concept of a genuine preference to select F over H at t_3.

Similarly, the concept of a genuine preference (henceforth the CGP) is the time-independent rule that a cyclical ordering violates. To the same extent that one may violate the law of noncontradiction in one's speech behavior by sequentially verbalizing contradictory beliefs without intentional operators, one may also violate the CGP in one's selection behavior by sequentially making contradictory – i.e. cyclical – pairwise comparisons. And just as we may conclude in the belief case that under these circumstances, the agent has no intelligible belief at all, similarly, in the preference case, we may conclude that the agent has no CGP at all. In rejecting the distinction between a genuine preference and the selection behavior that may or may not reveal it, the canonical language of preference thereby conflates the distinction between an abstract, time-independent rule and its physical, time-dependent application.

Moreover, the CGP is the rule violated by Clyde, who, you will recall, sacrificed psychological consistency in order to preserve transitivity in the canonical form of (T). What enabled Clyde to alter all his priorities from moment to moment was an inability to compare and recognize his present ordering as the same as or different from previous ones. And this inability stemmed from an absence of perceived enduring similarities between present and previous alternatives that would have motivated him to recall them to mind at present. Clyde might have had some volatile and transient preferences, but he would not have known it. For like Wallace, he would be incapable of inductive generalization over his experiences to the CGP. Only an abstract rule such as the CGP furnishes could have provided Clyde with a criterion of consistency for recognizing such similarities from moment to moment, and therefore for conforming his preferences to this criterion.

But *the criterion of consistency <u>any</u> genuine concept or rule provides is in the end always one and the same, i.e. that some concrete particular shall not exemplify both it and its negation at one and the same time and in one and the same respect.* That is: the criterion of consistency a genuine concept provides is a criterion of *logical* consistency. The psychological inconsistency of Clyde's preferences,

therefore, was an unavoidable consequence of their intertemporal logical inconsistency.

But logically inconsistent preferences are psychologically inconsistent as well. Recall that Winifred fell into psychological inconsistency only at t_3, when her selection of H over F produced a cyclical ordering. This selection ensured that there could be no subset of pairwise comparisons that contained at least one selected alternative in common. But this psychological inconsistency was the consequence of failing to apply consistently the CGP to the alternatives presented. Failure to apply this concept, and the consequent absence of a common selected alternative, also accounted for Wallace's learning problem, i.e. his difficulty in inductively generalizing to this concept. We have seen that in order to form and consistently apply such a concept to one's selections among alternatives, those selections must be psychologically consistent. But the selection of one alternative in common to two pairwise comparisons implies the consistent application of the concept of a most preferred alternative, i.e. of a genuine preference, on both trials. So an agent's selections are psychologically consistent if and only if they are intertemporally logically consistent.

Nonvacuous utility-maximization presupposes the subordination of (T) to the requirement of logical consistency expressed in the concept of a genuine preference. Suppose an agent in fact ranks F highest at t_1 if and only if she ranks F lowest at t_3; Cleopatra's swift and lethal disposal of her nightly suitors in Theophile Gautier's tale might illustrate such a case.[41] Here any particular suitor is most preferred at night if and only if he is least preferred the following morning. This sort of arrangement detaches Cleopatra's love life from the integrated continuity of the rest of her life. While she may integrate her practice of lethal one-night stands into her life, she cannot integrate any love relationships into it. For she fails to develop and experience the depth and intensity of emotion that accompany prolonged engagement with a romantic other. Her present feelings and responses will tend to harden into stereotypes, because they are of the kind elicited only by the earlier, formulaic stages of courtship and never by the later, more individuated ones she has chosen to forego. After many repetitions of an initially novel or exciting experience, these responses may tend to become gestural or perfunctory, or so highly refined that their meaning evaporates – not only because of their repetitive character, but more importantly because she has chosen to foreclose to those responses an open-ended future. In thus voluntarily constraining their object, duration, and course of development, Cleopatra constrains in advance the range of significance and consequences they can have for her,

[41] Theophile Gautier, "The Nights of Cleopatra," in *Mademoiselle de Maupin* (New York: Modern Library, 1949).

and so their potential interest for her. Her nightly courtships thus become an empty and dilatory ritual, practiced on an unsuspecting other with whom genuine emotional exchange is thereby precluded. Unless there exist some peculiarly constructed agents for whom such empty rituals can be regarded plausibly (and nonvacuously) as utility-maximizing, the conclusion seems obvious that Cleopatra's romantic arrangement renders her quest for utility-maximization self-defeating.

The same conclusion holds whenever FsG at t_n if and only if GsF at t_{n+1}, where the relevant units of time are, say, days or weeks, or even months. In all such cases, the utility of F at t_n is at least partially obviated by its disutility at t_{n+1} for an agent who is otherwise supposed to endure through time.[42] The reason why it would be rational for Cleopatra to maintain longer-term intertemporal logical consistency among her preferences is because, like all agents with the capacities both to recall the past and anticipate the future, something experienced as a source of utility at one point that later becomes a source of disutility becomes a source of greater disutility in light of its history. Far from being eradicated entirely in its earlier guise, it is later experienced for that very reason as a loss or disappointment. And for theoretically rational creatures who can anticipate this outcome at the outset, this expectation itself tends to diminish the original estimate of its earlier utility. This is a perfectly general argument that does not depend on the particulars of this example for its force. So the same conclusion derived from considering Clyde's, Winifred's, and Wallace's failure to preserve short-term logical consistency among their preferences applies here as well: Preserving longer-term logical consistency among our preferences is also necessary, in order to be able to intentionally rank alternatives in the first place. Nonvacuous utility-maximization presupposes conformity to the requirements of logical consistency.

It would seem, then, that the principle of maximizing utility under the Ramsey-Savage interpretation is, as argued, best understood as instantiating a more comprehensive principle of logical consistency in selection behavior, and modified accordingly. In this case, it is possible that there are other, equally contingent principles of rational action that are rational just in virtue

[42] But can't I prefer sleeping most in the late evening and least the following morning without at any time decreasing the utility of sleeping? Of course. But in so far as I do so, I do not actually *select* sleeping in the late evening nor waking the following morning at all; they *steal over* me. And in so far as I do actually prefer sleeping most in the late evening and least the following morning, it is not at t_n and t_{n+1} respectively that I do so, but rather from some rationally distanced perspective on my health habits that is time-independent of both. To say in what that rationally distanced perspective consists requires a conception of rationality that avoids the problem of funnel vision described in Chapter II.2.3. I try to develop that conception in Volume II.

of satisfying the requirements of that more comprehensive principle, even though they demand of the agent no utility-maximization in the nonvacuous sense. For example, the principle that, under certain circumstances, an agent is to achieve her ends with respect for tradition might satisfy this more comprehensive principle as well: If Agnes prefers the messenger-delivered, handwritten note on embossed stationary to the telephone at t_1, and the telephone to the e-mail at t_2, then logical consistency – not mere, defenseless transitivity – requires her to prefer the messenger-delivered, handwritten note on embossed stationary to the e-mail at t_3. Even though the e-mail may be more efficient in the nonvacuous sense, tradition and elegance triumph nevertheless. This is because, if the above arguments are sound, to preserve intertemporal logical consistency in action may be, but is not by definition, to maximize utility. Since the alternative that an agent must rank most highly in order to preserve at least one most preferred element from one pairwise comparison to the next temporally contiguous one will depend on her actual preferences and circumstances, whether the requirement of logical consistency is to be satisfied by maximizing utility or in some other way will depend equally on those circumstances.

This use of the concept of a genuine preference also has the important virtue of blocking the charge that the (suitably reformed) concept of utility maximization is vacuous. For if an agent's selection of alternatives on the third trial is logically inconsistent with her selections on the first and second, as (U) allows it to be, we cannot assume that she is merely changing her mind, as the principle of charity would have it. We now have some reason to suspect she might be *losing* it.

4. The Utility-Maximizing Ideal

That according to (U), all our behavior comes out rational raises questions, not only about the viability of utility theory as an explanatory and predictive theory of human behavior, but about (U) as a higher-level principle of interpretation. (U)'s higher-level status is often invoked to excuse its vacuity against some of the arguments presented here. It is claimed that vacuity is a necessary side effect of a principle that normatively enjoins us to interpret all the phenomena of human behavior in its terms, and that it is therefore exempt from the Popperian requirement of falsification that lower-level hypotheses derived from it must meet. But (U)'s purported normative status throws into relief the contingent values it expresses. I shall refer to these values conjointly as expressing the *utility-maximizing ideal*. Dissecting this ideal in some detail will show that the inconsistencies mentioned so far stem from flaws inherent in the basic conceptualization of (U), rather than in faulty and therefore corrigible applications of it.

The utility-maximizing ideal expresses the idea that the envisioned alterations in oneself or one's environment provide, as objects of value or

desire, the sole justification for performing the actions that effect them. In one sense, this idea is so obvious as to be tautological on the face of it: We act as we do in order to achieve the valued ends that motivate us. These ends may include the action itself as a valued object, in addition to its causally or conceptually distinct consequences. But this tautological interpretation of the idea is not the one that the utility-maximizing *ideal* expresses. (U) as a descriptive principle may be vacuous, but its prescriptive implications are not. The point is rather that the utility-maximizing ideal regards physical human action and its practical extensions – tools, machines, other people, etc. – as only derivatively valuable or significant; i.e. only in so far as they instrumentally promote its intended consequences.

To the utility theorist this point may not be sufficient to distinguish the utility-maximizing ideal from commonsense tautology. It is true that if, for example, behaving with dignity is Miles' intended end, then that is at least a conceptually distinct consequence of the physical actions he undertakes – forbearing from reciprocating insults, refraining from malicious gossip, comporting himself with poise in hostile situations, and so on. And it would be hard to think of any conceptually distinct action that is not instrumental at least to this degree.

To see what makes the utility-maximizing ideal meaningful and distinctive, we need to draw on the distinction made in Chapter III, Section 4.1, between structure and intention. We saw there that the relation between forbearing from reciprocating insults and behaving with dignity may be *structurally instrumental*, if the latter is a conceptually distinct consequence of the former. But forbearing from reciprocating insults is *intentionally instrumental* to behaving with dignity only if Miles' only reason for forbearing is to behave with dignity; i.e. only if there is no further value he attaches to forbearing beyond that derived from behaving with dignity. Thus what is at issue between the tautological and the meaningful interpretations of the utility-maximizing ideal is not the structural relation between physical action and its causally or conceptually distinct consequences. We can agree that there is always, or perhaps even necessarily, an instrumental element in this relation. The question is rather whether or not one is inclined to regard such an action favorably, independently of the instrumental value conferred on it by its promotion of those intended consequences.

It is not difficult to conceive of a person who does view forbearing from reciprocating insults favorably for additional reasons, besides her desire to behave with dignity. Forbearing may be valuable to Ruby because, due to her upbringing, she is naturally disposed to behave that way. Perhaps trading insults makes her feel social uncomfortable or ill at ease, or makes her feel ashamed of herself. Of course each of these feelings could be reformulated as intentional ends Ruby desires to achieve – to avoid social discomfort or self-condemnation, say – by forbearing from reciprocating insults. But by

hypothesis such formulations would be inaccurate, since they are not in fact intended consequences of her physical actions.[43] They are instead internalized dispositional constraints on her behavior that lead her to experience forbearing favorably *an sich*. Aside from wanting to behave with dignity, Ruby forbears because that is the most natural, comfortable, and acceptable way for her to act. Her physical actions are valuable to her as actions, aside from the consequences she intends to effect by performing them. This is the sort of person whose attitude toward action furnishes a counterexample to the utility-maximizing ideal.

The utility-maximizing ideal implicitly denies the theoretical interest of such a case. It fails to find anything of theoretical significance in physical human action in itself, independently of its structurally instrumental relation to its intended consequences. Of course the utility-maximizer recognizes the inescapability of the effects of conditioning or habituation on an agent's attitude toward her own behavior – for example, as the swimmer comes to experience swimming as intrinsically valuable, aside from its intended health benefits or competitive goals. But from the utility-maximizing perspective, this value is not really intrinsic. Either it must be interpreted as a utility the swimmer attempts instrumentally to maximize by swimming, or else it is theoretically irrelevant. The concept of physical human action as itself a source of noninstrumental, rational value or interest is meaningless on this view.[44]

Again it may seem that this is a merely logical consequence of an action's being described by its intended ends, whether it is physically basic (such as raising one's arm)[45] or complex (such as signaling a turn). But it does not follow from the fact that all actions are described in terms of their ends that the relation of the actual physical behavior to that end must be either

---

[43] For the same reason, it would be conceptually incoherent to regard these ends as intentionally instrumental means to the final end of expected utility-maximization. If Clarissa does not forbear from reciprocating insults because it ultimately makes her happy (in a nonvacuous sense), then she does not intend to secure her happiness by doing so.

[44] Richard Brandt states this view explicitly when he asserts that "in large part human behavior is an instrument, also. For we care about how people act mainly because of how their behavior affects other persons for good or ill, by disappointing their expectations, injuring them, and so on. For the most part, acts are important to us because of their consequences." (*A Theory of the Good and the Right* (Oxford: Clarendon Press, 1979), 196-7.

[45] See Arthur Danto, "Basic Actions," in Care and Landesman. Also see J. L. Austin, "A Plea for Excuses," in *Philosophical Papers*, Ed. J. O. Urmson and G. J. Warnock (New York: Oxford University Press, 1970), 175-204.

structurally or intentionally instrumental.[46] We have already seen in Chapter III, Section 4.1 that the relation of actual physical behavior to the end of a basic action such as raising one's arm is *structurally constitutive* if the physical behavior is invariably linked with that end, i.e. if the description of the one is equivalent to the description of the other. Here the physical behavior of raising one's arm is distinct from that of one's arm rising, although both might be nonintentional under certain circumstances (for example, I might reflexively and unintentionally raise my arm in response to the national anthem, or to a random synaptic firing. These cases are different from those in which my arm simply goes up, without any experienced connection to my sense of agency at all, as in hypnosis.). There is no *structurally instrumental* element in this relation if the end is not a conceptually distinct consequence of the physical behavior: The end of raising my arm is not a causally or conceptually distinct consequence of my physical behavior; it *is* my physical behavior. This description is satisfied by many basic physical actions. Therefore, more complex actions that include them include a structurally noninstrumental element as well.

Similarly, the relation of actual physical behavior to the end of a basic action is *intentionally constitutive* if an agent's intention to perform that physical behavior is inextricably linked to her intention to achieve that end; if *a fortiori*, the performance of that actual physical behavior is her end. There is no *intentionally instrumental* element in this relation, if she does not intend to perform the physical behavior in order to achieve the end. Again this description is satisfied by many basic physical actions. Therefore, all more complex actions that include them include an intentionally noninstrumental element as well. So the utility-maximizing ideal that views physical human action as theoretically interesting or valuable only because of its instrumental relation to the ends it enables us to achieve is not a tautological consequence of our concept of action as described by its ends. The explanatory origin of this distinctive stance lies elsewhere.[47]

The distinctiveness of the utility-maximizing ideal as a perspective on action illuminates some further implications of utility theory. To define rational action in terms of efficiency or utility-maximization means that the fractional ratio of the resources an agent expends in action to the number and importance of ends she achieves should be as small as possible, and the

[46] Although the single end interpretation (Chapter III, Section 2) shows that utility theory requires us vacuously to construe it that way.

[47] See Max Weber, *The Protestant Ethic and the Spirit of Capitalism*, Trans. Talcott Parsons (New York: Charles Scribner's Sons, 1958) and Albert O. Hirschman, *The Passions and the Interests* (Princeton: Princeton University Press, 1977).

smaller the ratio the more rational the action.[48] Thus that physical human action has only instrumental value conferred on it by the ends it promotes implies that the less of it we need to perform in order to achieve those ends, the better and more rational our behavior is. This is why, when constructing an ideal model of rational action, Neoclassical economists abstract from certain natural characteristics of actual human action that are viewed as flaws or constraints on efficiency: an agent's limited mobility, the time lag between desire and achievement, incomplete information regarding available resources for achieving her ends, limited foresight of the consequences of alternative courses of action, uncertainty with respect to their probability distribution, and so forth.[49] The view of such characteristics as limitations on an agent's efficiency implies that in the ideal limiting case, a rational agent should achieve her ends instantaneously, with no expenditure of time or energy whatsoever. This is the way to achieve the smallest possible fractional proportion of resources expended to ends achieved. So the limiting ideal of utility-maximization implied by (U) requires that an agent's adoption of an end physically and temporally coincide with its realization. That is, it describes a situation in which instrumental human action is eliminated entirely. The ideally rational action, then, is not really an action at all, but rather an atemporal set of instantaneous desire-satisfaction events, relative to which any instrumental effort at all is an unwelcome deferment of gratification.[50]

---

[48] See Tibor Scitovsky, *The Joyless Economy* (New York: Oxford University Press, 1977), 65.

[49] For critiques of this maneuver, see Ward Edwards' work on the existence of preferences for particular probabilities of winning at gambling over others (e.g. of a 4/8 over 6/8 probability of winning) independently of utility considerations, in "Experiments on Economic Decision-Making in Gambling Situations," *Econometrica 21* (1953), 349-350; "Probability-Preferences in Gambling," *op. cit.* Note 23; "Probability Preferences Among Bets with Differing Expected Values," *American Journal of Psychology 67* (1954), 56-67; "The Reliability of Probability Preferences," *American Journal of Psychology 67* (1954), 68-95; "The Theory of Decision-Making," *op. cit.* Note 31; Donald Davidson, Sidney Siegel, and Patrick Suppes, "Some Experiments and Related Theory," *op. cit.* Note 31; Herbert A. Simon, "A Behavioral Model of Rational Choice," *Quarterly Journal of Economics 69* (1955), 99-118; and "Rational Choice and the Structure of the Environment," *Psychological Review 63,* 2 (1956), 129-38); and more recently, the work of Amos Tversky and Daniel Kahneman, *op. cit.* Note 30. George Katona discusses these constraints in "Rational Behavior and Economic Behavior," *Psychological Review 60*, 5 (1953), 307-318.

[50] Michael Slote has described this (in conversation) as the "let there be light" model of human action.

Notice that this attitude extends to desired future events. For me to want at t_1 that the world be such that event E occurs at t_n, without any expenditure of instrumental effort or resources on my part in the interim, is to want now that the course of objective events conform itself to the desire I now have, whether for present or for future gratifications. It is also, by implication, to want at any future moment that that course of events conform itself with equal exactitude to the desires I have at that moment, whether for present or for future gratifications.[51] That is, it is to have a second-order desire about the manner and efficiency with which my first-order desires are satisfied, namely, that they are to be satisfied the moment I desire that they be satisfied, without any expenditure of effort on my part in the interim. And in the ideal case, this second-order desire itself is to be gratified instantaneously as well.

Thus the limiting ideal of utility-maximization implies that there is no order or sequence in which events should occur that is independent of when an agent desires their occurrence. Like the belief-desire model of motivation, the utility-maximizing model of rationality is *egocentric*, in that it describes a limiting ideal in which the time, place, and manner in which desired events occur are entirely dependent upon the agent's desires as to when, where, and how they should occur. Of course a egocentrist is sensitive to information regarding predictable causal sequence, the pace of natural processes, and so on, and defers the satisfaction of her desires accordingly: The egocentrist is, of necessity, ordinarily forced to be an opportunist. But even this much moderation is provisional. The egocentrist regards ritual, custom, and the natural order of things, like the natural characteristics of human action, as obstacles to the maximization of utility which it is the basic project of utility-maximizing rationality itself to overcome.[52]

---

[51] See Richard Brandt's discussion of the incoherence of this want in *A Theory of the Good and the Right*, Chapter XIII, Section I, 247-253 (*op. cit.* Note 44).

[52] For alternatives to this way of thinking, see, most recently, Jeremy Rifkind, *Time Wars* (New York: Henry Holt and Co., 1987), Chapter 8. A contrast is also provided by E. E. Evans-Pritchard's description of the Nuer of the Egyptian Sudan:

> [T]he Nuer have no expression equivalent to time in our language, and they cannot, therefore, as we can, speak of time as though it were something actual, which passes, can be wasted, can be saved, and so forth. I do not think that they ever experienced the same feeling of fighting against time or of having to coordinate activities with an abstract passage of time, *because their points of reference are mainly the activities themselves*, which are generally of a leisurely character. Events follow a logical order, but they are not controlled by an abstract system, *there being no autonomous points of reference to which activities have to conform with precision* (italics added).

(E. E. Evans-Pritchard, *The Nuer: A Description of the Modes of Livelihood and Political Institutions of a Nilotic People* (Oxford: Clarendon Press, 1940), 103; quoted in Staffan B. Linder, (New York: Columbia University Press, 1970), 19.

Intrinsic to the utility-maximizing ideal, then, is the tendency, found in more fully elaborated, decision-theoretic formulations of (U), to eliminate the natural characteristics of human action as independent variables. But it is equally evident in the three interpretations of utility sketched in Chapter III, Section 4, which eliminate human action as an independent variable by redefining every action as itself a case of utility-maximization.[53] Thus on the one hand, we never succeed in attaining the limiting ideal of utility-maximization because we must always perform some instrumental action, however minimal, in the service of our goals. But on the other, we invariably attain it by definition of having acted at all. In utility theory, human action is both crucial and irrelevant to full rationality. It is both identical with and dispensible to utility-maximization. Utility theory furnishes an ideal of rationality that is both everywhere and nowhere instantiated.

Notice, further, that these two possibilities paradoxically coexist. The description of the ideally rational action as omniscient, omnipotent, and instantaneously effective is of a kind of action that maximizes utility just by being performed. Similarly, the stipulation that actual human action reveals the agent's preferences, i.e. maximizes utility, implies that it, too, is of this kind. But actual human action is not, or at least rarely of this kind. If it is true that any human action must incur some opportunity costs, then the limiting ideal of egocentric utility-maximization is *in theory* unattainable. The obstacle to its attainment is not simply that we are not smart, quick, or resourceful enough. The obstacle is a conceptual one: We can *never* be smart, quick, or resourceful enough, no matter how much so we are, because we are not the immaterial but omnipotent will that the limiting egocentric ideal requires. That there is a distinction between what we intend and what previously exists, and between what exists and what we do, is implicit in the concept of a conscious and individuated agent. If there were no such distinctions, utility would be maximized – continually, universally, and necessarily. What exists would be logically equivalent to what was desired. Agency would be unnecessary. Thus it is the necessity of agency that constitutes the real obstacle to the attainment of this limiting egocentric ideal. It is inaccessible to us, not because we are not rational enough, but rather because we are, indeed, agents.

So consider further the case in which there are no such distinctions, and no such agents in the strict sense; i.e. the case in which all that exists is the continual, universal and necessary maximization of utility to which the

[53] As Gauthier puts it, "the economist does not define an individual's utilities, and then ask whether the individual seeks to maximize utilities so defined. Rather, he determines what the individual seeks to maximize, and then defines his utilities accordingly." ("Economic Rationality and Moral Side-Constraints," *Midwest Studies in Philosophy III: Studies in Ethical Theory* (Minneapolis: University of Minnesota Press, 1978), 76.

utility-maximizing ideal aspires. This is the limiting state of bliss for Max U., in which he does nothing and receives everything he wants, every minute, all the time. A perpetual, universal and necessary desire-satisfaction event eliminates all instrumental expenditures in its service, and therefore has no costs. Similarly, such an event eliminates all wants or lacks that it might have replenished, and therefore satisfies no desire. For as we have seen in Chapter II, within that value system a desire-satisfaction event loses its inherent value at the moment it occurs. Hence in this limiting, blissful state, neither costs of achieving it nor the experience of achieving it are available to confer value on it. This means that this event is not only *gratis*, but – in Max U.'s value system – therefore worthless; for no sacrifices have had to be made that might confer value on it, and no more overriding values of any other kind are available to take up the slack. In the limiting case, then, in which Max U. gets everything he wants at the exact moment he wants it, this not only costs him nothing but also means nothing to him.

Of course Max U. is only an abstraction; and the continual, universal and necessary maximization of utility in perpetual desire-satisfaction events is only an ideal. But it is worth meditating on the condition of actual human agents who, often for reasons beyond their control, asymptotically approach this ideal: who get virtually everything they want the minute they want it, without paying any significant costs of achievement, either before or after the fact; who are never required to strive or sacrifice in order to realize their goals, and who are never required to compensate for mistakes in their choice of goals or the strategies for achieving them; for whom, therefore, no such choice can ever have negative implications or consequences, and for which they therefore never need take responsibility. For such agents, it is all good, all the time, whatever "it" is; and because whatever "it" is is worthless and meaningless in the sense just described, it does not matter what "it" is in the end. The limiting ideal of perpetual utility-maximization robs agents of their abilities, robs goals of their value, and robs lives of their meaning, leaving in its wake victims who cannot understand why the instant gratification of their every wish increases their unhappiness and disconnectedness rather than alleviates them. To instill or cultivate this ideal in the young is effectively to destroy them. I flesh out some of its further psychological and social implications in Volume II, Chapter V.6.1 of this project.

5. Efficiency vs. Ethics

The argument so far has been that utility theory is either riddled with paradox, or else expresses a contingent value that is limited in its scope of application and subordinate in status to the requirements of logical consistency. Now, finally, I cash out the suggestion that utility theory and moral theory compete for our practical and theoretical allegiance.

Despite the conceptual and normative paradoxes of utility theory, it, like moral theory, performs a regulative function for its adherents. It determines their behavior by determining their interpretation of their experience – as does any explanatory theory the behavior of those who believe it. In this respect, utility theory, like, for example, Freudian theory, is entirely on a par with moral theory; and each is likely to determine their adherents' behavior differently. Suppose, for, example, that Oswald's subordinate Daphne complains loudly to him over what she considers to be a paltry salary increase. If Oswald is a Freudian, he may be more inclined to speculate on the subconscious motives and meanings of her behavior, and to respond to those assumed subconscious motives and meanings more than to the explicit ones. He may interpret Daphne's tantrum as evidence of an unresolved Electra complex rather than as frustration over her meager raise, and so help her work through her anger, rather than increase the size of her raise. But if Oswald is a utility theorist, he may be more inclined to assume that Daphne is acting in order to maximize her expected utility, try to discern the advantage her action promotes – to embarrass and ultimately unseat him, perhaps, and react to the promotion of that perceived advantage – say, by calling her bluff – rather than to the conventional or explicit meaning of her action. These possibilities are no different in status than that moral interpretation which understands Daphne's temper tantrum as an expression of moral outrage at the lack of respect for her job contribution her pittance of a pay raise expresses. Nor does this third interpretation differ in its capacity to motivate a distinct sort of response from Oswald: contrition, grudging respect, a promise to rethink the profit margin to which he aspires, etc. All of these possibilities suggest that moral theories and social theories both perform a regulative function, with roughly comparable success. When confronted by such a case we are often uncertain how to react. This may be because we are uncertain which theory of another person's behavior to believe.[54]

Moral theory and social theory are also comparable in the valuational status of the ideal type[55] in each respectively, and indeed of each kind of theory more generally. A theory may be value-laden in two ways. First, it may mention certain values within the value-conferring part of the theory itself, as, say, a theory of religion mentions what is good, right, and divine relative to particular religions. Second, it may itself express or promote certain values in its action-guiding part, whether or not it mentions any within the value-conferring part. For example, a theory of scientific method that promotes the values of deductive inference, intersubjective confirmation, and experimental

[54] I elaborate this point in Volume II, Chapter IX.7.

[55] in Weber's sense; see *The Theory of Social and Economic Organization*, Ed. Talcott Parsons (New York: Free Press, 1964), Chapter 1.1-2.

testing does this. I discuss this distinction at greater length in Chapter V.2, below.

Now it may seem that a moral theory like Kant's is value-laden in both of these respects, whereas as utility theory is value-laden in neither. First, the value-conferring part of Kant's moral theory mentions certain values, such as autonomy, freedom, trustworthiness, and beneficence. These are the values that, according to Kant, guide the behavior of a perfectly rational being. In this respect, Kant's concept of a perfectly rational being is comparable to what Weber calls an ideal type of *Wertrationalität*.[56] Second, the action-guiding part of Kant's moral theory promotes these values, in so far as it offers us a regulative ideal to which to aspire in our conduct.

By contrast, some would say that the concept of fully rational economic man in Neoclassical economics – roughly Weber's ideal type of *Zweckrationalität* – is value-laden in neither of these senses, or at least is so to a much lesser degree. First, it does not claim to mention any values; on the contrary. (U) purports to describe value-neutrally an agent who pursues whatever values she has efficiently. Second, since (U) purports to express the basic principle of an explanatory social theory, it appears to promote only the value of computational rationality itself. These two reasons alone have been invoked to justify ascribing to the concept of utility-maximization that special, logically necessary and therefore value-neutral status in the concept of rational action which, it is claimed, even Kant himself conceded.

I do not believe Kant conceded this, although I shall not defend this belief here. But I have tried to show in the preceding sections of this discussion that the concept of utility-maximization or efficiency itself is a fully contingent value an agent may rationally choose to reject in favor of some competing alternative. Among the competing alternatives from which an agent may choose are to be found the values that load a moral theory such as Kant's: autonomy, respect, beneficence, trustworthiness, and so on. That Kant's moral theory, and the values that define it, compete with utility theory and the values that define *it*, is uncontroversial: Values like trustworthiness versus efficiency, duty versus self-interest, beneficence versus personal satisfaction continually vie for importance and for our attention, and present us with familiar conflicts both in theory and in practice. The two theories that respectively contain these values are not significantly different in regulative or valuational status, as some social theorists have sought to argue.[57] If this is

---

[56] *ibid.*

[57] Weber represents the reasoning of the *Zweckrationalität* advocate about *Wertrationalität* particularly clearly: "From the [*Zweckrationalität*] point of view, ... absolute values are always irrational. Indeed, the more the value to which action is oriented is elevated to the status of an absolute value, the more 'irrational' in this sense the corresponding action is. For, the more unconditionally the actor devotes himself to this value for its

true, and if utility theory regulates the behavior of its adherents in the ways I have tried to suggest, then there is no intrinsic difference in prescriptive status between moral theories like Kant's and a highly developed social theory like that of Neoclassical economics. These two theories express contingent, competing, equally regulative theories of value. So a moral theory like Kant's is to be distinguished from a social theory like Neoclassical economics solely by its content, i.e. by the values it both mentions and promotes.

Now each of these two kinds of theories purports to subsume the other as a special case. Utility theory treats adherence to principles of the sort found in Kant's moral theory as a special case of instrumental, utility-maximizing behavior under special conditions in which the agent is one of many equally rational, powerful, and self-interested agents with equal access to limited resources that benefit each. Under these circumstances, adherence to moral principles is rationally justified as a means of avoiding Prisoner's Dilemma-type situations when they threaten,[58] but not otherwise. By contrast, moral theories like Kant's treat adherence to utility-maximizing principles as rational only under the constraint that it not violate overriding and fundamental moral principle: that one's efficient actions be willable as a universal law, for example; i.e. that they satisfy Kant's principle of rational consistency. Under these circumstances, utility-maximization is rationally justified in so far as it promotes or at least does not undermine overriding moral principle, but not otherwise.[59] Thus each of these two kinds of theories does not merely compete for our commitment to the values they express. They also compete for regulative and theoretical superiority within our conceptual scheme. If the preceding extended argument, that utility-maximization is only one contingent value among many, is well-taken, then it is also an argument for the more comprehensive principle to which we have seen it must be subordinated as a special case, namely the principle of logical consistency that governs philosophical reasoning. I have not argued that Kant's universalization principle of rational consistency is identical to this, although I argue elsewhere that Kant thought it was. But that Kant's principle presupposes it, while canonical utility theory does not, is clear.

---

own sake, to pure sentiment or beauty, to absolute goodness or devotion to duty, the less is he influenced by considerations of the consequences of his actions" (Weber, *ibid*., p. 117).

[58] See, for example, John Rawls's *A Theory of Justice* (Cambridge, Mass.: Harvard University, 1971), Part I; also David Gauthier, "The Incompleat Egoist: From Rational Choice to Moral Theory," *The Tanner Lectures* (Stanford University, 1983), esp. Part II: "What Should an Egoist Do?"

[59] Here, see Rawls, *ibid*., Part III.

Chapter V. A Refutation of Anscombe's Thesis

In Chapter I I framed this project as a defense of a Kantian conception of the self against the prevailing Humean paradigm. In her influential mid-twentieth century article, "Modern Moral Philosophy,"[1] G. E. M. Anscombe takes a different view. She sees the Kantian tradition as historically dominant. She argues that we should now abandon the Kantian law conception of moral philosophy as an anachronistic relic of a religious, divine-command sensibility without which moral philosophy itself, traditionally understood, is impossible. In her opinion, Kant's moral theory is a *deontological* theory that requires a religious motivational foundation that is now lacking. She characterizes the more contemporary *consequentialist* trend in modern moral philosophy as equally impotent without an adequate philosophy of psychology to support it, and calls for the development of one. Call this division of normative moral theories into consequentialist and deontological *Anscombe's thesis*.

I do not accept Anscombe's thesis. Nor do I accept her interpretation of the history of moral philosophy and its vicissitudes. In fact, I would reverse the historical order she proposes. I believe, rather, that the anachronistic paradigm that has gripped Anglo-American moral philosophy (and indeed philosophy and the social sciences more generally) for the last two and a half centuries is the one which Anscombe describes as "consequentialist" – that is, that model of reasoning instrumentally about moral action which has its origins in Hobbes' *Leviathan* and finds its fullest and most detailed expression in Book II of Hume's *Treatise*. I would argue that it is now time to move on to the historically more modern, "deontological" alternative that prescribes moral action in accordance with the dictates of conscience rather than with its probabilistically anticipated outcomes. I see the "consequentialist" paradigm as an ideological relic of that ancient and primitive style of magical thinking (which has its place, although not in academic philosophy) that supposes one's ability to causally determine, through individual symbolic physical acts, the unforeseeable course of the universe; and that conflates control of external consequences with self-determination and personal divinity.

I argue below that this paradigm in any case has no practical application to any actual normative theory developed, and that it is too rigid and schematic ever to do so. Nevertheless, Anscombe's thesis has generated a virtual cottage industry among philosophers who, on the one hand, choose not to address normative issues; but who, on the other, similarly choose not to address the perennial metaphysical and metapsychological issues of traditional metaethics. Anscombe's thesis has stimulated a seemingly endless

[1] *Philosophy 33* (1958), pp. 1-19.

debate that, by generalizing and arguing about what consequentialist-*type* and deontological-*type* theories require or do not require or imply or presuppose, enables one to steer clear both of overwhelmingly complex casuistical issues and also of the deeper metaphysical value commitments these issues presuppose. My aim in this chapter is not to dampen this debate,[2] but merely to establish a rationale for those who might prefer not to join it. Briefly stated, my rationale is that Anscombe's thesis is false, and that the distinction between Humean and Kantian metaethical views is the philosophically fundamental one for the twenty-first century.

Anscombe contrasts "consequentialist" theories in general with a Kantian theory in particular as an example of what she calls a "deontological" theory. Whereas a Kantian theory is a certain kind of normative view, Anscombe's consequentialist/ deontological distinction sorts normative theories into those which assign primary value to the effects of actions – Utilitarianism, Aristotelianism, Marxism, and Perfectionism would all be putative examples; and those which purport not to, such as Kantianism, Intuitionism, or Moral Sentiment Theory. I do not think Anscombe's thesis is defensible. But it is important to say why, in order to clear the ground for the more appropriate metaethical distinctions I go on to defend. I argue here that given the content and structure of any normative theory we are likely to find palatable, there is no way of uniquely breaking down that theory into either consequentialist or deontological elements. In fact, once we examine the actual structure of any such theory more closely, we see that it can be classified in either way arbitrarily. So if we ignore the metaethical pronouncements often made by Anscombeans, we find that the consequentialist/ deontological distinction contributes nothing of consequence to an understanding of moral philosophy.

There are basically two reasons for this. First, what we mean by the terms endemic to the consequentialist/ deontological distinction have no unique references to particular states of affairs in actual cases of moral decision-making. Hence we may justify any such concrete moral decision by reference to typically consequentialist or deontological reasoning indifferently. Second, scrutiny of actual and viable normative theories reveals a much finer-grained structure than the cosequentialist/deontological taxonomy can capture. And

[2]As one prominent moral philosopher put it, the nice thing about Anscombe's thesis is that it keeps a lot of people busy so that the rest of us can get on with the hard tasks of figuring out how to behave and how to make the world a better place. And when this chapter was first excerpted for publication in article form, one up-and-coming Anscombean commented about it that my analysis was very likely right but had no relevance for his work, since he had no interest in actual normative theories, whether past or future, and moreover thought it presumptuous to compete with the Great Thinkers by proposing one. I asked him what he thought the consequentialist/ deontological distinction purported to refer to, if not actual normative theories. Idealized *possible* normative theories, he replied.

it is this structure, rather than simple attention to consequences or principles, that determines practical moral decision-making. We would thus do better to develop the richer vocabulary of causes and constituents, goals and effects, states and events (mental, social, or physical). When we do so, we see that all so-called consequentialist theories in fact presuppose the Humean conception of the self. So in the end, Anscombe's thesis is irrelevant at the normative level of actual moral reasoning, whereas at the metaethical level it crudely schematizes two opposing types of dummy theory, neither of which is of use to normative moral philosophers who seek in their work practicable solutions to actual moral, social, or political problems.

Section 1 begins by distinguishing two uses to which the consequentialist/ deontological distinction can be put. First, it can be applied to the construction of a theory of what is morally valuable, i.e. good or right. Call this the *value-theoretic* part of a normative theory.[3] Second, it can be applied to the construction of practical principles that guide deliberation. Call this the *practical* part of the normative theory. These two aspects of a moral theory are mutually independent, and normative theories need not be uniformly consequentialist or deontological with respect to both of these parts. I consider briefly the normative theories of Kant and Aristotle as examples of views that are mixed in different ways with respect to these two parts. Once we make this distinction between the value-theoretic and the practical parts of a normative theory, no such theory can be characterized as either uniquely consequentialist or uniquely deontological. Section 2 compares the practical parts of a purportedly consequentialist normative theory – namely Classical Utilitarianism – with the practical parts of a purportedly deontological one – namely Ross's Intuitionism. The comparison shows that we may submit the action any such theory prescribes to either characterization arbitrarily. Section 3 examines the value-theoretic part of a normative theory. It argues that the *content* of such a theory can, in turn, be distinguished from its *structure*; and that value-theoretic content is interchangeable between consequentialist and deontological theories, while there are no inherent structural differences between them. So the consequentialist/ deontological is as superficial to the value-theoretic part of a normative theory as it is to the practical part. Section 4 argues that the

[3]Thus I use the term "moral value" (or worth) to refer broadly to that which is morally evaluated. This includes both what William Frankena calls "moral value" (i.e. moral goodness and badness) and what he describes as "moral obligatoriness or rightness" (*Ethics*, Second Edition (Englewood Cliffs, N.J.: Prentice-Hall, 1973), p. 62). It is not clear how to characterize our moral attitudes to that which we deem right, if we cannot say that we value it, just as we value that which is good. My distinction between the value-theoretic and the practical parts of a normative theory resembles Holly Smith's distinction between moral theories as such and their uses as practical action-guides in "Making Moral Decisions," *Nous XXII*, 1 (March 1988), pp. 89-108.

consequentialist/ deontological distinction between normative theories is rather to be located in intensional metaethical attitudes proponents of these theories take toward them. But these attitudes suggest a different distinction which cuts across the consequentialist/ deontological one, namely between those normative theories which are *person-regarding* and those which are *theory-regarding*. Anscombe's thesis, then, finally implies a metaphilosophical claim about the moral psychology of philosophers rather than one about the adequacy of normative theories. Section 5 concludes by showing that all so-called consequentialist moral theories in fact depend for their models of motivation and rationality on the Humean conception of the self.

1. Values and Practice

1.1. Value Theory

The first of the two ways in which the consequentialist/ deontological distinction can be used is *value-theoretically*. Here the distinction is formulated in such a way as to distinguish between two approaches to the construction of a moral theory (thus I speak of "value-theoretic uses," "value-theoretic senses," as well as "value theories" simpliciter, according to context). On this view, a consequentialist theory[4] is one that begins by defining the good, i.e. the state(s) of affairs that is (are) claimed to have intrinsic value, e.g. happiness, pleasure, or perfection. The right, or morally obligatory, is then characterized as that which is conducive to the good.[5] The right may include, for example,

[4]I.e. that which Rawls and Frankena call a "teleological" theory, and which Brandt calls a "result" theory (see John Rawls, *A Theory of Justice* (Cambridge, Mass.; Harvard University Press, 1971), p. 24; William Frankena, *ibid.*, pp. 14-17; and Richard Brandt, *Ethical Theory* (Englewood Cliffs, N.J.: Prentice-Hall, 1959), p. 354. So-called "consequentialist" theories are actually only one possible kind of teleological theory, since, not all final ends of action are necessarily causal consequences of action.

[5]Frankena distinguishes between the morally and the nonmorally good on the basis of the subjects this predicate applies to. On his view, only persons, groups of persons, and elements of personality (such as motives, intentions, emotions, and dispositions) may be morally good, whereas practically anything, including physical objects, experiences, and forms of government may be nonmorally good. The two bases for this distinction are (1) ordinary usage; and (2) the reasons for which we make the judgment of goodness, which are not further elucidated. My own linguistic intuitions disincline me to accept this distinction. But more important, I find no distinction in the range of reasons for which I might make such judgments that would lead me to accept it. Why should not happiness be viewed as a moral good, just as is virtue? Why cannot democracy be judged to be just as much a moral good as rational beings as ends in themselves? Now Frankena does argue that "it does not make sense to call [things like experiences or forms of government] morally good or bad, unless we mean that it is morally right or wrong to pursue them." (*Ethics, ibid.*, p. 62). But neither would we think virtue or rationality were moral goods unless we thought it was morally right to pursue

actions the results of which are characterized as good, or institutions the effects of which are so characterized. In either case the good is then described as having priority over the right in the sense that the actions, institutions, or states of affairs that conduce to it derive their moral value from this fact alone, and all such acts and institutions are to be evaluated according to this criterion.

A deontological theory, by contrast, is one that defines the right independently of the good. It is argued that the moral value of an action or institution deemed right by the theory depends on other properties of it besides its consequences.[6] Such other properties might include, in the case of an action, how it was decided, or whether it conforms to certain more general moral prescriptions intuitively known to be valid. In the case of an institution, the relevant valuable or right-making property might include having evolved in a certain way, or expressing certain central interests or values of the community it is intended to serve, for example, as the institutions in Rawls's well-ordered society express the value of respect for persons.

In both cases, a common characteristic of the value-theoretic use of this distinction is that normative theories as so classified do not issue immediate directives to action. A consequentialist theory like utilitarianism that defines the good as, for example, the greatest sum of happiness on the whole for all sentient beings, and the right as that which is maximally conducive to this, does not prescribe any particular action or kind of action that it would therefore be right to perform in order to realize this end under particular circumstances. This purely value-theoretic part of utilitarianism leaves open the possibility that no individual action might be conducive to happiness; or that only institutions, and not individual actions, might promote this end. Similarly, a deontological theory that defines the right as that which conforms to certain general moral injunctions intuitively known to be true, such as keeping promises, does not enjoin us to perform any *particular* actions under a given set of circumstances. In both cases, the respective kinds of value theory provide different normative accounts of what is valuable or worthwhile, relative to which particular actions or institutions can be assessed.

In addition, value theories by themselves abjure specification of how, or in what sense, their particular normative values are to be promoted. They describe a purely conceptual or methodological priority relation between

---

them. Nor could we think certain individuals were morally good if we simultaneously denied that they were worthy of emulation. But if this is the criterion, then experiences, objects and forms of government can be moral goods after all. Happiness is a moral good for the Utilitarian, just as the Koran is for the Muslim, and just as Socialism is for the Marxist. I therefore ignore this distinction.

[6]I use the term "right" to cover duties, obligations, and recommendations indifferently for the time being. The importance of further distinguishing between uses of this word is taken up in Section 2.

what the theory stipulates to be good and what it stipulates to be right – without, however, specifying how the conceptually prior value is to be realized: causally or constitutively. To claim, for example, that justice is the highest good and that the good has priority over the right implies that those actions, institutions, or states of affairs are right which promote justice, and only insofar as they do so. But justice can be promoted causally, for example by effecting dispositions to just behavior in oneself and others, or constitutively, by acting justly or participating in just institutions oneself. Terms like "promotes," "furthers," or "realizes" are neutral between these two possibilities, and the value-theoretic part of a moral theory does not explicitly commit itself to either. Often the choice is made at the practical level, where the action-guiding directives prescribe how the value is to be promoted under particular circumstances. But this matter of value-theoretic policy is not made explicit *as a policy* at the practical level. Typically we just assume, when a value theory announces itself as consequentialist, that its conceptually prior value is to be promoted causally and instrumentally, whereas the value espoused by a deontological theory is to be promoted constitutively. But these assumptions are mistaken, for they suppose that a choice between these two possibilities is precisely what distinguishes value theories as consequentialist or deontological. In Section 3 of this chapter I show that the failure of such value theories to commit themselves explicitly one way or the other is better explained by the fact that any acceptable value theory must include both causal and constitutive relations, and so that no such distinction can be made.

Examples of *purely* value-theoretic normative theories that contain no practical parts are Rawls's theory of justice as originally elaborated in his book of that title, and Plato's theory of justice in the *Republic*. In both cases we are presented with an elaborated conception of the just society and a rationale for adopting it as a social ideal. But in neither case are we given any guidelines for bridging the gap between this ideal and our actual social condition. By contrast, Marx's social ideal of the truly human society is buttressed by an immediate call to revolutionary activity on the part of the proletariat in the service of this ideal. To be sure, the directive to overthrow the bourgeois system of exploitation through revolution does not specify the prescribed actions in the degree of detail one might like. But the degree of abstractness with which a prescribed action is described does not prevent it from being a practical prescription. Rawls' and Plato's normative theories contain no such prescriptions at all.

1.2. Practical Decision-Making

A second application of the consequentialist/ deontological distinction is therefore to the formulation of these prescriptions or directives to action. Call this the *practical* use of this distinction. Here the distinction differentiates between two different methods for deciding what to do. The consequentialist

method directs us to decide what to do by evaluating the expected outcomes of alternative available actions with reference to some wanted or valued state of affairs, and to perform that action most conducive to it. The deontological method bids us invoke other criteria for making this decision: It may, for example, direct us to perform that action the maxim of which can be consistently willed as a universal law of nature; or to perform that action we intuitively know to be right. In either case, the *method* for deciding what to do does not supply normative *value criteria* for deciding what to do. Rather, it supplies a particular model of moral deliberation.

Writers who observe the consequentialist/ deontological taxonomy have not been sensitive to the further distinction between its value-theoretic and practical uses. Frankena, for example, begins by characterizing a teleological moral theory as one that "says that the basic or ultimate criterion or standard of what is morally right, wrong, obligatory, etc., is the nonmoral value that is brought into being," and concludes a few paragraphs later that "in order to know whether something is right, ought to be done, or is morally good, one must first know what is good in the nonmoral sense *and* whether the thing in question promotes or is intended to promote what is good in this sense."[7] That is, he thinks it follows from the independent and prior characterization of the good typical of a consequentialist or teleological theory in the value-theoretic sense that the practical decisions of a person who accepts this theory must take a consequentialist cast; that the person must decide what to do by evaluating the outcomes of her actions with a view to promoting the good that is value-theoretically characterized.

Similarly, Brandt, in explaining Ross's deontological or formalist theory of prima facie obligations, criticizes it as incomplete on the grounds that "it is not possible to infer, from the principles he explicitly states, what is our duty in a particular situation ... even ... when it is known which act would maximize the welfare of sentient beings ... [and] with full factual information at our disposal, because he does not give us the second-order (much less third-order) principles necessary for determining our obligation overall, when prima facie obligations conflict."[8] Again, the suggestion is that a complete deontological theory implies a method for deriving practical directives for action that are as deontological in character as the substantive theory of value itself.

1.3. Kant's Mixed Theory

But there is no reason why consequentialist value theories need to be linked with practical consequentialist decision-making methods, nor why deontological value theories need to be linked with practical deontological

[7]Frankena, *op. cit.* Note 3, pp. 14-15.
[8]Brandt, *op. cit.* Note 4, pp. 393-394.

decision-making methods in the way these writers assume. One may, for example, adopt a consequentialist value theory that defines the good as welfare for all sentient beings, and the right as those actions that promote this; but not decide what action to perform on the basis of whether its *actual consequences* are in fact likely maximally to effect this goal. Instead one may use this initial characterization of the good and the right to develop a list of types of action that, under specified circumstances, would ideally *constitute* maximizing the welfare of all sentient beings (such as: when driving in the country drive slowly and observe wild animal crossing signs; when in city parks, feed the pigeons; when making more than four times a subsistence-level salary per year, distribute at least an eighth to relief funds; etc.), and perform these actions when the circumstances obtain, irrespective of their actual expected outcomes. Thus the consequentialist value-theoretic conception of the good would be linked to a practical deontological account of right action. The result would be a theory of moral action that attempts noncausally to realize a conception of the good by acting in the way the constituents of this conception itself seem to require, rather than in the way its causal achievement seems to require.

Kant's normative theory can be understood to have such a form. Although his conception of the highest good includes happiness, defined as a pleasant feeling, the supreme condition of the highest good and its most important component is virtue, i.e. the worthiness to be happy.[9] But the concept of virtue is then explained to be that of a will – the good will – all of whose maxims conform to the moral law (2C, Ac. 32-33), i.e. all of whose resolutions to action could serve as universal laws.[10] Kant then maintains that to require that an agent's maxim, or resolution to act, be capable of serving as a universal law is the same as to require that the maxim be such as could serve as law in a kingdom of ends, i.e. of rational beings: these are just two different formulations of one and the same categorical imperative, the supreme principle of morality (G, Ac. 434, 436). Thus the highest good includes a will whose maxims, or resolutions to act, could effectively operate as law in a community of beings, each of whose will conforms to the same conditions. Now the concept of a good will, all of whose maxims satisfy this requirement can only be an ideal toward which human beings strive (2C, Ac. 32). And indeed Kant claims that we can only seek the highest good in the

[9]Immanuel Kant, *The Critique of Practical Reason,* trans. Lewis White Beck (New York, N.Y.: Bobbs-Merrill, 1956), Ac. 110. Henceforth all Academy Edition references to this work are parenthecized in the text, preceded by "2C." For purposes of this chapter I confine myself to what Kant says, leaving aside the question of why, and whether he ought to have said it. I take up these matters in greater detail elsewhere.

[10]Immanuel Kant, *Groundwork of the Metaphysic of Morals*, trans. H. J. Paton (New York, N.Y.: Harper Torchbooks, 1964), Ac. 402. Henceforth all Academy Edition references to this work are parenthecized in the text, preceded by "G."

concept of an intelligible or supersensible world of fully rational beings (2C, Ac. 21, 68, 114-115, 119), which is identical to the concept of a kingdom of ends (G, 452, 458).

Thus Kant directs us to adopt as a final end an ideal of action. This ideal is part of an ideal end-state, i.e. the law-governed kingdom of ends, which is in turn one characterization of the supreme moral requirement we must *actually* aim to satisfy in all our actions. This requirement on action is claimed in turn to be constitutive of the end-state, i.e. the highest good, which is achieved by satisfying it.

So when Kant enjoins us to regard ourselves qua rational beings as making laws in a kingdom of ends which is possible through freedom of the will, and then argues that "morality consists in the relation of all action to the making of laws whereby alone a kingdom of ends is possible (G, Ac. 434)," he can be interpreted as making two claims. First, the kingdom of ends is indeed an intrinsic good. For Kant its value is not contingent on any considerations extrinsic to that conception itself (G, Ac. 428; 2C, Ac. 87). Nor does the full characterization of the kingdom of ends invoke moral notions of what is in some further sense good or right. Moreover, like other purely value-theoretic consequentialist theories, actions are defined as right just insofar as they promote the realization of this conception: keeping promises, for example; or developing one's talents and capacities. But of course Kant does not *practically* prescribe the performance of those actions whose *consequences* might *causally effect* this conception. Rather, we are to perform those actions which themselves *constitutively promote* this conception, irrespective of their causal consequences. And we know which ones those are by submitting the maxims of our actions to the consistent universalization procedure described by the first formulation of the categorical imperative – a clearly deontological method of practical decision-making. In Section 3 of this chapter I argue that there is a sense in which any value-theoretic consequentialist theory must adopt some such brand of practical deontological decision-making method.

1.4. Aristotle's Mixed Theory

Conversely, one may adopt a deontological value theory in conjunction with a practical consequentialist decision-making method. One may develop particular criteria of right action that do not depend on the good they can be expected to cause, but rather, for example, on what is required of a morally virtuous individual. Such a theory might prescribe as right those actions which such an individual would perform (courage in the face of danger, for example; or fulfilling one's responsibilities, or honesty), irrespective of the ends thereby effected. But in deciding what to do, one might adopt the practical consequentialist method of choosing to perform those actions the expected outcomes of which best promote the end of becoming such a morally virtuous individual, or of performing those actions such an individual would

perform. Here the result would be a value theory of right action the practical prescriptions of which enjoin those actions that maximally effect the performance of the morally required actions, rather than those morally required actions themselves.

In some cases, the prescribed action might then be one the description of which coincides with the favored description of the morally required action. For example, if the value theory makes telling the truth morally right, the practically prescribed action might consist in uttering a particular set of true sentences under certain circumstances. Here the desired consequence of the action – telling the truth – would be identical with the performance of the action itself, and therefore with that morally right action prescribed by the theory.[11] Under other circumstances, however, the goal of telling the truth might necessitate a period of prolonged psychological self-scrutiny and intensive behavioral conditioning designed to negatively reinforce the tendency to lie compulsively. Or it might necessitate the uttering of a set of sentences some of which are true and some of which merely express favorable or unfavorable emotions and therefore have no truth value, together with those unambiguous behavioral attitudes that are often crucial to the distinction between uttering true sentences and telling the truth. In these cases the practically prescribed actions would not be identical with those specified as morally right by the value theory.

There is much in Aristotle's moral theory to suggest such a reading. Aristotle's claim that the good for human beings consists in the performance of that function proper to them, i.e. "an activity of the soul in conformity with excellence or virtue,"[12] is fleshed out in Books II, III-IX of the *Nicomachean Ethics* to refer to the development and practice of the moral virtues, guided by practical wisdom and intelligence. Aristotle's conception of the good is therefore not defined independently of a prior conception of morally right action.[13] This is evident from Aristotle's remark in Book II that the virtue or excellence of human beings (i.e. moral virtue) is what makes a person good and able to perform his function well (1106a15-23). To say that moral virtue makes a person good, that the final good is the exercise of moral virtue is to suggest that the final good to be aimed at is one's own moral goodness or excellence as expressed by one's character and one's actions – a moral ideal of right conduct that already has been defined by the deontological criterion of performing our proper human function.

---

[11]Brandt recognizes this, but without explicating its implications for the consequentialist/deontological taxonomy. *Op. cit.* Note 4, p. 354, n. 2.

[12]Aristotle, *Nicomachean Ethics*, trans. Terence Irwin (Indianapolis: Hackett, 1985), 1097b22-1098a17. Also see 1144a6-9. Henceforth citations are parenthecized in the text.

[13]Here I ignore for the sake of argument the controversies surrounding the correct interpretation of Book X relative to the *Nicomachean Ethics* as a whole. In fact, I ignore Book X and the problems it raises altogether.

Indeed, the role of the cultivation and practice of the moral virtues in Aristotle's theory lends plausibility to the view that the notion of the good as so defined plays the expected teleological role only in the most superficial sense. For although Aristotle assures us on the one hand that virtuous conduct is that which is truly constitutive of happiness (1098b30, 1099b15-1100a5, 1100b11-1123, 1101a12-21, *passim*), the final good at which all actions aim (1095a19, 1097a34-1079b20, *passim*), he takes great care to emphasize at the same time the fact that a truly virtuous individual performs noble acts for their own sake and not for the rewards they will bring (1120a22-25, 1140b6, 1116b20-30, 1144a18-20, *passim*). A virtuous person continues to act virtuously when bad fortune has crushed her chances of supreme bliss (1100b17-1101a14),[14] in the face of death in certain forms (1115a32-35), and without regard to the pleasurable or painful consequences of action as such (1140b11-20) – as we would indeed expect from a person whose actions were the consequence of traits of character deeply instilled by habituation.

Thus moral virtue is not prescribed simply as that means best suited to achieving the highest good of happiness. On the contrary, moral virtue is that brand of conduct which constitutes the ideal of happiness itself. Morally virtuous conduct for Aristotle both defines the final good and causes its achievement. And it generates a list of morally obligatory actions (e.g. courage, generosity, temperance, etc.) that are determined by the noninstrumental consideration of what our proper human function consists in, rather than the actual consequences they can be expected to effect.

Practically considered, Aristotle's normative theory is a consequentialist one. In deciding what to do, we are to choose those actions the expected outcomes of which promote the development in us of the moral virtues, and hence the highest good. Because truly virtuous action for Aristotle issues from deeply inculcated dispositions of character, the full description of the action we practically ought to perform coincides with that of the morally right action only in the limiting case, in which we have already achieved the ideal of moral virtue. Otherwise, Aristotle enjoins us to practice performing through imitation those actions which truly virtuous individuals perform in a virtuous way (1103b5-23, 1105a25-1105b8), to aim at the mean, or moderation, in cultivating virtuous dispositions to action and feeling (1106b5-7, 15), to avoid that extreme which is most opposed to the temperate feeling or action in question (1109a30-35), and to be particularly circumspect when considering actions to which we feel naturally inclined, or which afford us personal satisfaction (1109b2-12). Thus we are to act in ways that causally develop in us the capacity for performing the morally required actions that characterize the

[14]Here Aristotle distinguishes clearly between happiness, of which noble action performed for its own sake is constitutive, and supreme bliss or contentment as a state of mind consequent on good fortune.

truly virtuous individual – a clearly consequentialist method of practical decision-making. In Section 3, below, I argue that there is a sense in which any value-theoretic deontological theory must adopt such a practical consequentialist decision-making method.

These readings may be thought to go against the grain of the received interpretations of Kant and Aristotle. Kant's theory at first glance resembles a purely deontological one because it practically prescribes the performance of certain actions without regard to their causal outcomes. But the value-theoretic end that they nevertheless promote – rational nature as an end in itself – is what determines their moral worth. Similarly, Aristotle's theory at first glance resembles a purely consequentialist one because it practically prescribes those actions which causally effect the highest good. But the highest good is then value-theoretically characterized as virtue of character and action, the worth of which is not in fact contingent upon their effecting some further end. For as we have seen, these actions are to be performed even when the prospects of contentment, pleasure, and indeed continued life itself are dim.

So the identification of Kant's and Aristotle's normative theories as respectively deontological and consequentialist is plausible only if we ignore the distinction between the value-theoretic and the practical aspects of these theories. I next show in Sections 2 – 4 that when this second distinction is taken into account, no normative moral theory can be adequately described as either consequentialist or deontological. Anscombe's thesis propounds a distinction without a difference.

2. Practice Re-examined

2.1. Classical Utilitarianism

Now to examine consequentialist and deontological theories respectively, considered in their capacity as practical decision-making methods. Classical Utilitarianism is often taken to be the paradigm consequentialist theory, and Anscombeans will be quick to argue that it remains so even when the suggested value-theoretic / practical distinction is recognized. Value-theoretically, the good is independently defined as the greatest possible sum of happiness, and morally right actions are defined as just those that promote this end. Practically, we cannot know the actual consequences of our actions with one hundred percent certainty, nor even their objective probabilities. Rather than concluding from this with Moore that therefore we can never know which of our actions are right,[15] Utilitarianism commonly prescribes as morally right just those actions that we can reasonable expect to promote the

[15]G. E. Moore, *Principia Ethica* (Cambridge: Cambridge University Press, 1968), pp. 149-50.

very same end, i.e. the greatest sum of happiness – a clearly consequentialist practical decision-making method. Or so it is claimed.

But this claim is false. This practical prescription describes a deontological decision-making method that evaluates the moral rightness of actions independently of their consequences. For it is not whether some action actually promotes the greatest amount of happiness that determines its rightness, but rather whether it can be reasonably expected to do so. This means that the action is right even in case, contrary to reasonable expectation, it does not do so. This is as it should be. For surely a Utilitarian would hesitate to withdraw the appellation "morally right" in that one anomalous case out of a hundred in which retrospective information demonstrated that the action had not had best consequences after all. It would hardly be practically helpful to be able to assign moral rightness to actions only retrospectively, on the basis of the consequences they actually happen to have had, for this would furnish no guidance at all as to what we ought to do next. Indeed, given that we can never know the totality of the consequences of any action, this would make it impossible to assign moral value to actions with any degree of certainty at all.

This difficulty cannot be remedied by providing a Utilitarian theory of excuses, according to which actions at least could be characterized as praiseworthy or blameworthy in the event that we could never know whether they were objectively right or wrong.[16] According to such a theory, we could then prescribe or proscribe actions based on their degree of moral culpability, rather than on their rightness or wrongness. But from the point of view of practical deliberation, this just locates the deontological feature of putatively consequentialist deliberation at a different point. For now we must base our decision of what to do not on a consideration of whether it can be expected to promote the greatest happiness or not, but rather on that of whether it can be expected to elicit praise or blame. And we can be wrong about this as well. Nor does the fact that the expected outcome – praise or blame – can be internalized as a motivation within the agent solve the problem. For in fact we are often motivated to act in just that way which we anticipate and hope will allow us to keep peace with our consciences – and find that we were mistaken. Here, too, it may be only a retrospective examination of the actual consequences of the action that reveals whether we are morally culpable or not, i.e. whether we ought to have been praised or blamed for performing it.

These considerations may not seem enough to transform the apparently consequentialist decision-making method into a deontological one. For the action's moral worth is still conferred by the good end defined by the theory.

---

[16]See Richard B. Brandt, "A Utilitarian Theory of Excuses," *The Philosophical Review LXXVII,* 3 (1969), pp. 337-61. Brandt has identified himself (in conversation) as a deontologist.

That is, the action still appears to have no moral worth independently of its relation to this end, regardless of whether the relation is one of cause and effect or one of cause to expected effect. But this appearance is misleading. To perform an action because one expects it to have certain consequences, and to think that the performance of this act will effect those consequences, is to intend to bring about those consequences. To then claim that an act is morally right because of one's intentions in performing it and not because of what actually happens as a result of performing it is to make one's intentions, and not the action's consequences, the criterion of moral rightness. Thus the practical prescriptions of Classical Utilitarianism are deontological in structure because they make the moral rightness of an action contingent on considerations other than its consequences, i.e. on its intended consequences. That this holds true – indeed, must hold true – for any other purportedly "purely" consequentialist normative theory can easily be seen.

2.2. Ross's Intuitionism

Next consider a supposedly pure deontological theory such as Ross's. The central value-theoretic claim is that we have certain general *prima facie* duties that rest on morally significant circumstances of action, and which are known immediately and intuitively to be true. They include, for example, duties of fidelity (such as keeping promises or telling the truth), of reparation (such as punishment), of gratitude (such as repaying a favor), of self-improvement, and so on.[17] But because these *prima facie* duties may conflict under certain circumstances, and because we cannot be certain which should take priority, our practical duty under particular circumstances is not similarly self-evident. Here the best we can do is consider the situation carefully, weigh the alternatives, reflect on our moral intuitions, and finally act in conformity with that considered opinion as to what act is probably our duty to the best of our understanding. And in this case it does appear that the practical method of deciding what act to perform is as deontological in character as the value theory from which it derives. In both bases, actions are prescribed as morally right without reference to their consequences.

But appearances are misleading in this case as well. A theory that characterizes as morally right the fulfillment of some duty independently of its consequences at the same time makes the actual fulfillment of that duty the criterion of rightness, rather than any expectations or intentions one may have had in the particular action one actually performed. And then the rightness of the action actually performed depends on its consequences after all. If the action does not have the effect of fulfilling the prescribed duty, it was wrong;

[17]W. D. Ross, *The Right and the Good* (Oxford: Clarendon Press, 1968), p. 21. Henceforth all citations to this work will be parenthecized in the text.

and if it did, it was right. As Ross argues, in discussing the example of keeping a promise by returning a book through the mail,

> nonattainment of the result proves the insufficiency of the means – however carelessly I pack or dispatch the book, if it comes to hand I have done my duty, and ... if the book does not come to hand I have not done my duty. Success and failure are the only text, and a sufficient test, of the performance of duty (45).[18]

Again this is as it should be. A deontological theory that practically enjoined us only to attempt to keep promises and repay our debts to others could be followed successfully even though moral duties were never fulfilled. Indeed, such a theory would not even require us to adopt as a goal of action the fulfillment of these duties. We would be obligated only to try. But mere moral tryings cannot be the subject of moral prescription, for they need never enter into the description of any actual actions we perform. My trying to mail the book may consist in little more than a rebellious stirring of will that makes my actual act of throwing the book into the fireplace less than effortless or conflict-free. Here I could honestly say that I tried to mail the book and failed (because my effort of will was not strong enough). Thus such a theory would not prescribe moral actions at all, but rather moral motivation. And because good intentions are not the sort of thing we can immediately will ourselves to have, we would then be morally obligated to undertake the actions that would effect this change in character, rather than to fulfill the duties that the theory prescribes.[19]

So the practical prescriptions of a purportedly "pure" deontological theory are consequentialist in structure because they bid the performance of only those actions the actual outcome of which is the morally right action as specified by the theory. That this holds equally true for any deontological theory that practically prescribes certain kinds of action as morally right is easily seen.

2.3. Consequences, Intrinsic Value and Moral Beliefs

But consider the following objection to this argument.[20] Making the actual fulfillment of a moral duty an end to which particular actions are means does not suffice to transform practical deontological prescriptions into consequentialist ones, for the end in question is not defined in the way a consequentialist theory requires. A consequentialist theory, it might be said, does not evaluate an action merely by the positive character of its

[18]In general the discussion of pages 30-36 support this point, Ross's intentions notwithstanding.
[19]Ross recognizes this. See *ibid.*, p. 405.
[20]I owe this objection to Richard Brandt and Allan Gibbard.

consequences, but rather by how much intrinsic value it produces. Let us define *intrinsic value* as follows:

> X has intrinsic value =df. X would be rationally or fittingly desired for its own sake, independently of
> (1) one's moral beliefs;
> (2) its actual or believed consequences.

Think of clause (1) as the *Millian condition*, following Mill's definition of higher pleasures as those which are, among other things, desired independently of one's moral beliefs about what one ought to desire.[21] And think of clause (2) as the *Kantian condition*, following Kant's stipulation of that sole source of intrinsic value as having it independent of its consequences (G, Ac. 399-400). Take pleasure as the most uncontroversial example of an end that satisfies both conditions. The claim is then that a concern with consequences as such fails to turn a deontological view like Ross's into a consequentialist one, because such a view neither does nor can claim that the prescribed actions are worthwhile because of the intrinsic value of their consequences.

We may begin by conceding that no such purportedly deontological view does claim this, passing directly to the question of whether it should. I now answer the objection by showing that either it should, or else there is no such thing as intrinsic value. Consider the definition. Clause (2) is prima facie unproblematic. Clause (1) is important, because we have moral beliefs about what we ought to do. If any of these beliefs figure in our conception of an intrinsically valuable end, then that end itself at least partially consists in some characterization of what we ought to do. In that case the conformity of deontological prescriptions to the consequentialist canons of intrinsic value is straightforward. So stipulating the independence of intrinsic value from our moral beliefs is important for maintaining the distinction between consequentialist and deontological decision-making methods.

But it is difficult to produce an example of intrinsic value that is not dependent on our moral beliefs. Pleasure would not seem to be a good example of this. In order for pleasure to be an intrinsic good, we must believe that it is at least permissible to seek pleasure. This in turn implies that if we desire pleasure, other things equal, we ought to seek it. But we identify this as a rational "ought" only because we believe that it is rational to satisfy our instinctive desires, other things equal. However, this belief is a moral one, grounded in the norms of Hellenic culture. On this general view, we behave most morally when we give full expression to our natural human capacities: for abstract thought, for self-determination, and for pleasurable experiences of

[21]John Stuart Mill, *Utilitarianism*, Ed. George Sher (Cambridge: Hackett Publishing Co., 1979), p. 8

certain kinds. To claim, as Aristotle does (1097b23), that the goodness of human beings consists in performing their proper human function implies that we morally ought to do that which most fully expresses our humanity. We in the West assume that this must give a prominent – if not dominant – role to the pursuit of pleasure.

Consider an opposing, but equally plausible set of beliefs about what it is rational to do; call it the Advaita Vedantic view.[22] On the Advaita Vedantic view, the highest good is objective knowledge. However, to achieve this requires, not the full expression of human capacities as an end in itself, but rather their eventual transcendence. Abstract thought is criticized for reducing the richness of objective reality to manageable but solipsistic human categories, and true self-determination is seen as incompatible with the indulgence and gratification of our biological human desires. The pursuit of pleasure draws human beings even further into a world of illusion, ignorance and self-seeking because it limits our comprehension of reality to that which is consonant with our pursuit of sensory self-gratification. Hence it reinforces the illusion of individual ego-consciousness. Genuinely objective and nonillusory knowledge, according to the Advaita Vedantic view, can be achieved through ascetic practices, meditation, and detachment from the pleasures of the senses, i.e. through the renunciation of those sensory and psychological supports that sustain the illusion of the individual self. Hence not only does the Advaita Vedantic view deny that sensory pleasure is an intrinsic good that it is permissible to seek. It maintains that, on the contrary, sensory pleasure is a positive impediment to good that one ought strenuously to avoid. By contrast to the Hellenic view, which suggests that sensory pleasure is good because it expresses a human capacity, the Advaita Vedantic view maintains that pleasure is bad for precisely the same reason. On the Advaita Vedantic view, the pursuit of pleasure hinders that surrender and transcendence of individual selfhood that is a necessary condition of achieving objective knowledge.

Thus the conviction that pleasure is an intrinsic good depends upon moral beliefs about the value of expressing human capacities and satisfying human needs. Ultimately, it depends upon moral beliefs about the value of the individual self that these capacities and needs uniquely define, and so violates clause (1) – the Millian condition – of the definition of intrinsic value. So it seems that we must look elsewhere for some good that satisfies the above definition of intrinsic value, such that it can be rationally desired without our

[22]The brief sketch in this paragraph of course does not do justice to the depth and complexity of Vedanta philosophy as actually elaborated in such texts as *The Upanishad*s, *Bhagavad Gita*, *Yoga Sutras*, or Shankaracharya's commentary on the *Brama Sutras*. See the Bibliography for suggestions. Any reasonable translation of the *Bhagavad Gita* would be a good place to start; that by Swami Prabhavananda and Christopher Isherwood (New York: Mentor, 1972) is one of the most accessible.

believing that we morally ought, under certain circumstances, to pursue it. This enterprise seems unpromising. Or, we can relativize our judgments of intrinsic value to our moral beliefs, in which case the assimilation of deontologism is, as I have suggested, straightforward: The morally prescribed action is intrinsically valuable and, as in consequentialist theories, it confers moral worth on those actions necessary to realize it.

Ross's theory may be pursued instructively as an illustration of this point. Ross believes that the highest intrinsic value is a moral good, i.e. virtue (134, 155). After virtue comes happiness or pleasure (136-38) and knowledge (138-40). Virtue is defined as having good motives and performing good actions, i.e. goodness of character (134, 155-56). Good motives are in turn characterized as, among other things, acting from a sense of duty, i.e. being motivated to fulfill our duties (134). But our duties include not only things like fulfilling promises and the like, but also cultivating virtuous motives such as benevolence and sympathy in ourselves, i.e. the "duties of self-improvement" (21, 24, 160-61). To have a good character and so to be virtuous is to perform actions motivated in this way (155-60). Now Ross already has claimed that actions are right only if they succeed in producing the desired effects. And now we learn that the desired effects include the production of virtue – in addition to other intrinsic goods such as pleasure (or happiness):

> When we think of an act as right we think that either something good or some pleasure for another will be brought into being. When we consider ourselves bound, for instance, to fulfill a promise, ... [or] when we consider the other main types of duty – the duties of reparation, of gratitude, of justice, of beneficence, of self-improvement – we find that in the thought of any of these there is involved the thought that what the dutiful act is the origination of is either an objective good or a pleasure (or source of pleasure) for someone else (162; also see 134).

Some of Ross' views undergo metamorphoses in his later *Foundations of Ethics*[23], but that this is not one of them is clear from the following passage:

> An action will be completely good only if it manifests the whole range of motivation by which an ideally good man would be affected in the circumstances, a sensitiveness to every result for good or for evil that the act is foreseen as likely to have, as well as to any special *prima facie* obligations or disobligations that may be involved; and only if it manifests sensitiveness to all these considerations in their right proportions. But if the agent is responsible to all the morally relevant considerations in their right proportions, he will in fact do the right act. Thus no action will have the utmost moral excellence which an action in the circumstances can have, unless it is also the right action.[24]

---

[23](Oxford: Clarendon Press, 1939).

[24]*Ibid*., p. 309. I am indebted to Richard Brandt for bringing this passage to my attention.

The consequentialism of Ross's purportedly "pure" deontological theory is evident. A deontological theory that failed to have such implications would be one that claimed there was literally nothing to be gained by performing morally obligatory actions. This would exemplify a dummy deontological theory, in that its only function would be to serve as the bull's eye for consequentialists at target practice.

2.4. Prescriptive Indeterminacy

That we may apply the same reasoning in each of the two cases – the practical prescriptions of a "consequentialist" theory such as Classical Utilitarianism and the practical prescriptions of a "deontological" theory such as Ross' intuitionism – to the other is easily demonstrated. If the practical prescription to perform that action which can be expected to maximize happiness is deontological in character, we can just as easily argue that we must then in fact perform that action, or string of actions, which has the prescribed action as a consequence, since we will not always be able to perform that action which can be expected to maximize happiness directly. Hence the apparently consequentialist prescription, shown to be deontological, is in fact consequentialist in structure after all. Similarly, if the practical prescription to fulfill what we believe to be the moral requirements of right action is actually consequentialist in character, we can just as easily show that we must then in fact perform that action which can be expected to have the fulfillment of what we believe to be the morally required action as a consequence, since we cannot know with certainty the consequences of our actions before we perform them. So the apparently deontological prescription, shown to be consequentialist, is deontological in structure after all. Each of these arguments respectively can then be repeatedly reiterated for the conclusion to deontological or consequentialist structure respectively.

From this possibility the suspicion rapidly and justifiably develops that the practical prescriptions of consequentialist and deontological normative theories are themselves neither essentially consequentialist nor essentially deontological in structure. They can be formulated in either way, depending on what aspect of actually carrying them out we choose to emphasize. Call this the *prescriptive indeterminacy thesis*. The prescriptive indeterminacy thesis calls attention to the fact that it is true both that prior actions may need to be performed in order to achieve the performance of the prescribed one; and also that we can only choose actions on the basis of the outcome we can reasonably expect them to have – even when the outcome we want is the performance of the prescribed action itself. This is hardly an original observation:

> The maxim: "ignore the consequences of actions" and the other: "Judge actions by their consequences and make these the criterion of right and good" are both alike maxims of the abstract Understanding. The consequences, as the shape proper to the action and immanent within it,

> exhibit nothing but its nature and are simply the action itself; therefore the action can neither disavow nor ignore them. On the other hand, however, among the consequences there is also comprised something interposed from without and introduced by chance, and this is quite unrelated to the nature of the action itself.[25]

Anscombeans who actually study the historical record of *de facto* normative theories that moral and political philosophers have worked to elaborate might be persuaded of the inadequacy of the consequentialist/ deontological distinction to shed light on the practical prescriptions of any such theory.

3. Value Theory Re-examined

Next let us consider more closely the value-theoretic parts of normative theories. Here Anscombe's thesis raises two questions. First, is there any intrinsic difference in content that distinguishes consequentialist from deontological theories? And second, is there any intrinsic difference in their structures? An Anscombean of course would answer both questions affirmatively. I propose to answer both negatively. In this section I turn to the first question, leaving the second for Section 3.3, below.

3.1. Interchangeability

It may seem evident that there is a radical difference in the kind of content appropriate to consequentialist and deontological theories respectively. Here the basic issue on which the distinction turns is whether a moral theory is constructed so as to ascribe primary value to some end the realization of which serves as the criterion for evaluating the moral worth of actions or institutions that promote it; or whether it ascribes primary value to these actions or institutions themselves, independently of their outcomes. In the first case the end in question is commonly described as "good," and that which promotes it as "right." In the second case, the actions or institutions are held to be right on other grounds, and not just as means to some further end. But once again we will see that this distinction is not sufficient to distinguish between two normative value theories described as consequentialist and deontological respectively, for anything that can count as good in this sense can also be right, and anything that is right in this sense can also be good. So whether the right or the good is to have priority is of no importance for the substance of one's favored normative theory.

What confers moral value on whatever in the theory has worth or value? The consequentialist may claim that the end confers value on the actions and institutions that promote it, but that nothing further confers value on the end itself; it simply has intrinsic worth. We can describe this latter type of value as

---

[25]G. W. F. Hegel, *The Philosophy of Right,* trans. T. M. Knox (New York, N.Y.: 1975), paragraph 118, note.

primitive, meaning that the state of affairs in question is claimed to have intrinsic worth or value that is not dependent on its relation to any further end or condition. So, for example, Aristotle's concept of pleasure, although not instrumental to any further end, would not have primitive value in this sense, because it depends on other conditions – most notably, virtuous action – to confer value on it. Call final ends that have value in this sense *carriers of primitive value*, or CPVs.

Properties of CPVs can be cited in virtue of which the CPV has value. The utilitarian, for example, can point to the fact that happiness is something all human beings strive to achieve; the perfectionist can cite the fact that the final state of human perfection represents the full development and exercise of human capacities. But in neither case is this to supply some further condition or end that confers instrumental value on the ends in question. It is merely to explicate the relevant properties of these ends themselves that make them CPVs. Call these characteristics the *value-conferring properties* of CPVs.

Now the consequentialist's claim that the final end is the CPV has varying degrees of persuasiveness, depending on the final end involved. Moral theories that posit happiness, human flourishing, or survival as their final end can adduce the claim of primitive value somewhat more plausibly, perhaps, than those that posit pleasure or aesthetic appreciation. Those that posit riches, power, or security seem to hold considerably less title to this claim. Let us suppose that the *metaphysical structure* of some state of affairs specifies it as either a state or an event, and more specifically as a physical or a mental state, and as an activity or action, or an occurrence. Then we can see that among these theories, the plausibility of the claim of primitive value does not depend on the final end's being a mental state rather than an activity, or a physical state rather than an event. Happiness is as plausible a candidate for a consequentialist's value-theoretic final good as is the exercise of the human capacity for self-government; survival is as good a candidate as the achievement of ultimate self-knowledge. CPVs, then, must be distinguished by their content and not by their metaphysical structures.

The deontologist may answer the question of what confers worth or value on that in the theory which has value in much the same way as the consequentialist did with respect to the final good. The deontologist may begin by claiming that actions that fulfill moral duties, or fair democratic political institutions are also CPVs: They are inherently right and do not derive their worth from any further end or condition to which they are instrumental. It is nevertheless compatible with this claim for the deontologist then to go on to explain that the moral worth of fulfilling one's duties derives from its morally significant characteristics, as in Ross's theory (138), or from the fact that fulfilling one's duties expresses rational human nature, as in Kant's. Similarly, it might be argued that the morally important property of fair democratic political institutions is that these are institutions to which any

participant would explicitly agree upon careful reflection, or which would be chosen under certain intuitively acceptable ideal conditions. Again the value-conferring properties are not further, independent ends or conditions that fulfilling moral duties or democratic political institutions are intended to effect. Other, more efficient ways of expressing rational human nature would not displace the moral importance of fulfilling one's duties nor would other matters on which people would rationally agree displace the moral importance of fair democratic political institutions. To cite these properties is not to confer moral worth on right action or just institutions only instrumentally, any more than to cite the fact that all human beings strive for happiness is to make the worth of happiness instrumental to the further end of having all human beings strive for it. To cite these properties is rather to explicate what it is about these actions and institutions themselves that make them valuable. Thus deontological value theories have CPVs just as do consequentialist value theories.

Once again the plausibility of the deontologist's claim depends largely on what is value-theoretically asserted to be morally right. Fulfilling certain duties is a plausible candidate; as might be experiencing emotions such as guilt, remorse, shame, or resentment under certain appropriate circumstances; as might be, as well, social and political institutions that respect the privacy and freedom of its citizens. Less persuasive as CPVs might be, for example, consistently altruistic behavior; or feeling repentance for one's sins, or continuing political and social disequilibrium. Again the important point is that deontological prescriptions to bring about states of affairs perceived as inherently and self-evidently valuable need not be confined to morally obligatory actions. Once again that which is prescribed as right may as well be an activity as an emotion, an event as a state. What ought to be the case is neutral between these possibilities, and again it seems that CPVs must be distinguished by the their content and not their metaphysical structures.

But this then implies that any activity, mental or physical state, or event that can be a valued end relative to a consequentialist value theory can be, with respect to its metaphysical structure, the subject of a deontological value theory and vice versa. To experience happiness under the appropriate circumstances and to experience resentment under the appropriate circumstances are both states we can strive to experience as an end as well as states of which it makes sense to say we ought to experience. Hence both are states that can be constitutive of the consequentialist's final end as well as morally right on independent grounds. To express fully our human talents and to fulfill our obligations are equally activities that it might be good to perform as well as activities of which it makes sense to say we ought to perform them. Hence both are activities that can be constitutive of the final end as well as morally right. The achievement of universal suffrage and political reform are both events it might be a good thing to have occur as well

as events of which it equally makes sense to say they ought to occur. Hence both are events that can be constitutive of the final end as well as morally right. These examples merely illustrate the point that normative theories cannot be value-theoretically differentiated according to what I have called the metaphysical structure of their carriers of primitive value.

This is not to claim that all CPVs are interchangeable between any two consequentialist and deontological theories. A deontological theory such as Ross's that is couched in the stronger terminology of what is not only right but morally obligatory would intuitively rule out certain CPVs commonly associated with consequentialism. For example, it might be morally right to feel happy about certain things or under certain circumstances, but one would be hard put to find circumstances under which it would be morally obligatory to be happy. But of course the language of duty or moral obligation rules out certain deontological CPVs as well: Helping others is clearly the right thing to do under certain circumstances, but many would argue that the meaning of the word "obligatory" is such that it is never morally obligatory to do so. Similarly, a consequentialist that claimed of its final end that it was not only intrinsically valuable, but also the highest good, as Moore's Ideal Utilitarianism does, would rule out certain CPVs associated with both deontological and consequentialist theories of certain kinds. Thus we might be entitled to say that to feel remorse at the commission of a crime is intrinsically valuable as an expression of moral character; but it can hardly be described as part of the highest good. For it cannot be part of the highest good to have committed the crime in the first place. Similarly, a social Darwinist might plausibly claim that survival is inherently good, whereas the claim that it is the highest good would be considerably less persuasive.

However, it is nevertheless likely that for any CPV that is value-theoretically attached as a final end to a consequentialist theory, a plausible deontological theory could be constructed to which it would attach as the subject of deontological prescription; and that for any CPV value-theoretically attached to a deontological theory as the subject of deontological prescription, a plausible consequentialist theory could be constructed to which it would attach as a final end. Call this the *interchangeability thesis*. One example of the interchangeability thesis might include friendship and aesthetic experience as CPVs in Moore's ideal utilitarianism. Both of these could be easily prescribed as activities in which we morally ought to participate within the relevant deontological theory. Another example might be Rawls's two principles of justice, expressed in the institutions of a well-ordered society as CPVs in his deontological theory of justice. These, similarly, might well find a place as intrinsic goods in a consequentialist theory of social change. So carriers of primitive value may not be interchangeable in the strong sense that any one such carrier might occupy the relevant slot in any indifferently consequentialist or deontological normative theory. But they are value-

theoretically interchangeable in that it is the specific content of the normative theory, and not its consequentialist or deontological classification, that determines the suitability of any particular CPV to that theory.

So just as CPVs must be distinguished by their content and not their metaphysical structures within consequentialist and deontological theories respectively, CPVs in turn serve to distinguish among normative theories by their content and not by the consequentialist or deontological structure to which they are value-theoretically attached. So there is nothing in the value-theoretic content of CPVs that serves to distinguish normative theories into consequentialist or deontological.

3.2. Metaethical Convention

– Nothing, Anscombeans might complain, beside the convention metaethicists have stipulated in order to differentiate between types of normative theory; and the interchangeability thesis fails to respect this convention. Certainly we can use the words "right" and "good" to refer to anything we like. But the fact is that there exists an accepted metaethical practice of describing the most highly valued state of affairs within one's normative theory as "right" or "good" according to whether it is an action (or set of actions constituting an institution) or an end-state respectively. This is the rationale for the consequentialist/ deontological distinction at the metaethical level. Certainly the convention could have been different. But it isn't, and that alone is reason to abide by it.[26]

But this convention is not nearly as settled as all that. The distinction between actions and end-states is not clearer than those further distinctions it is intended to buttress. We have already seen that friendship counts as an end-state – rather than a relationship divisible into a set of actions – in Moore's Ideal Utilitarianism; and the full development of human capacities and talents as an end-state – rather than a set of actions – in perfectionism; whereas Rawls's well-ordered society counts as a set of actions or institutions – rather than an end-state. If normative ethicists make no rigorous distinction between actions and end-states, clearly they do not and cannot use the terms "right" and "good" in ways that would reflect this rigor – as indeed the examples already cited confirm. So the existing practice is considerably more diverse than the above complaint would have us believe. Although there are, of course, particular normative theories that do take this distinction with varying degrees of seriousness, there is no such convention at the normative level – regardless of the metaethical claims Anscombeans are often are inclined to make.

Certainly there might arise such a convention. We could fix a canonical use of the word "good" to denote only mental or physical states that involved

[26]I owe this objection to Allan Gibbard.

no actions – for example, thoughts and feelings, bodily states, particular distributions of resources, and so on. Similarly, we might stipulate the denotation of the term "right" to refer only to actions and sets of actions as that concept is understood in action theory. According to this convention, such things as happiness, shame, economic equality, knowledge, and physical fitness might be good on different normative theories. Neither friendship, human flourishing, or workers' control over the means of production could be good in this rigorous sense. These things instead would have to be designated as right, as would fulfilling – but not having fulfilled – one's duties, research, virtuous activity, and engaging in sex, sports, or other pleasurable activities. We would then have to say that, for example, virtuous activity was morally obligatory or right regardless of its consequences, as might be research, sports, or workers' control of the means of production; or that these were perhaps right only insofar as they resulted in happiness, knowledge, physical fitness, or economic equality respectively, and not otherwise. This certainly would be a very odd and counterintuitive convention.

So there is good reason for the existing heterogeneity of practice among normative ethicists with respect to what can be described as "good" or "right." It is that an interest in constructing a viable normative theory precludes the sacrifices of organization, content, and intuitive plausibility that strict adherence to the convention would require. The point can be generalized. We could, if we wanted, take the consequentialist/ deontological distinction as seriously as its more enthusiastic Anscombeans would like. But the resulting normative theories would be practically irrelevant and intellectually uninteresting. More on this in the following section.

3.3. Structural Equivalence

Now I turn to the purported structural differences between consequentialist and deontological theories. All normative theories contain the following basic elements:

(1) *Activit(ies)*, i.e. actions, institutions, or practices;
(2) *Final ends*, i.e. goals, objectives, or purposes;
(3) *Value-conferring propert(ies) of (2)*, i.e. those properties that we adduce to explain the value of the find end(s) of the theory.

We represent the basic and general structural relationships among (1), (2), and (3) as follows:

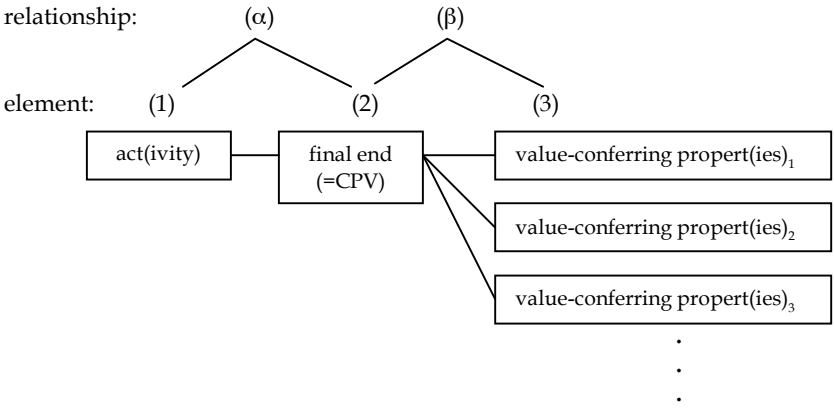


Figure 4. Structural Relationships among Basic Elements of a Normative Value Theory

Figure 3 is a schematic representation of the fact that in any normative value theory, there is an end to be achieved that is taken to have moral worth (2) and actions, sets of actions, or programs of action that are prescribed to achieve it (1). In addition, there are properties of that end (3) that, when enumerated, explain why that end is morally worthwhile or valuable.

Different normative value theories tend to construe the relationships (α) and (β) between these elements (1), (2) and (3) differently. Utilitarianism, for example, makes a sharp distinction between the action (1) and the final end it is intended to promote (2); whereas, as we have already seen in Section 2.2, a theory of moral obligation such as Ross's makes the prescribed actions (1) themselves the final end (2). Similarly, Perfectionism throws into sharp relief the value-conferring property of that end (3), namely that human potential is thereby fully developed and exercised; whereas Moore's Ideal Utilitarianism makes aesthetic experience an intrinsically valuable end apparently independently of any further properties it may be presumed to have. Here the final end (2) as such is identical with its value-conferring properties (3).

In Section 2 I observed that normative value theories do not uniquely specify their internal structural relationships merely by using terminology such as "promotes," "conduces to," "furthers," "realizes," or "makes possible."[27] This is because all these terms are neutral between causal and constitutive relationships, and between the actions to be performed and the values stipulated by the theory that confers moral worth on these actions. We assume that if a theory identifies itself as consequentialist, relationship (α) is essential causal, and so that the terms just listed are to be understood causally or

[27]This last is Kant's locution.

instrumentally. If the theory identifies itself as deontological, on the other hand, we assume that they are to be interpreted constitutively, so that relationship (α) is one of identity. Thus consequentialist value theories are thought to be distinguishable from deontological ones in virtue of the ways in which each construes the structural relationships (α) and (β) between elements (1), (2), and (3), although the major conflict concerns how (α) is to be construed. I first explicate in detail the structural properties that are assumed to distinguish consequentialist value theories from deontological ones. I then argue that these properties do nothing of the kind. Call this the *structural equivalence thesis*.

(i) In a *consequentialist value theory*, relationship (α) is usually described as (a) provisional, and/or (b) instrumental, and/or (c) causal.

> (a) Actions, institutions or practices (1) have only *provisional* value if the moral worth of performing or engaging in them is contingent upon their promoting the final end (2) specified by the theory. If they do not serve this end, they do not have moral value.
>
> (b) These activities promote their final end *instrumentally* if they are the means, medium, or instrument through which this end is achieved.
>
> (c) They promote their final ends *causally* if, roughly, they constitute a discrete set of physical conditions that produces a second discrete set of independently identifiable physical conditions, i.e. the end in question.[28]

(ii) In a *deontological value theory*, on the other hand, relationship (α) is typically characterized as (a) constitutive, (b) noninstrumental, and (c) noncausal.

> (a) Actions, institutions, or practices (1) have *constitutive* value if they are themselves the final end (2), or parts of the final end, which the value theory stipulates. This is, presumably, what is meant by saying that deontologically prescribed actions have intrinsic worth independent of their consequences. Thus in a deontological value theory element (1) is identical with element (2): the actions are "ends in themselves."
>
> (b) That relationship (α) is constitutive of the final end implies that it is *noninstrumental*: The action is not a means or instrument through which the carrier of primitive value is achieved. Rather it is itself such a carrier.

---

[28]This sketchy characterization is intended to reflect the view that causally related events must be physically separable. Cf. Jaegwon Kim, "Noncausal Connections," *Nous 8* (1974), pp. 41-52; Richard Brandt and Jaegwon Kim, "Wants as Explanations of Actions," *The Journal of Philosophy LX* (1963), pp. 425-35; Alvin Goldman, *A Theory of Human Action* (Englewood Cliffs, N.J.: Prentice-Hall, 1970); A. I. Melden, *Free Action* (London: Routledge & Kegan Paul, 1961).

(c) Finally, if an action under a certain description is identified as the CPV, it does *not cause* that carrier to occur. It can be said to promote that carrier only in some weaker sense in which it perhaps expresses, exemplifies, or actualizes it.

(iii) *Consequentialist and deontological value theories* tend to agree in their characterization of relationship (β) as noncausal, noninstrumental, and nonprovisional; and also (a) value-conferring, (b) explanatory, and (c) ascriptive:

(a) The relationship between elements (2) and (3) is *value-conferring* if it is the having of these propert(ies) that confers value on the final end (2) in question.

(b) The relationship is *explanatory* if adducing these properties explains why the final end has primitive value.

(c) The relationship is *ascriptive* if these properties can be ascribed to the final end as properties of it.

Thus consequentialist value theories make relationship (α) causal and (β) noncausal, whereas deontological ones make both relationships (α) and (β) noncausal.

3.3.1. Metaphysical Indistinguishability

Now to argue directly for the structural equivalence thesis, i.e. that these supposed structural distinctions between consequentialist and deontological value theories are largely illusory. First, note that according to the description of CPVs as those practices, states, or events that are claimed to be intrinsically valuable, those properties of CPVs which confer value on their carriers ((3)-type elements) are themselves CPVs, in both consequentialist and deontological theories. Thus, for example, the morally significant circumstances on which intrinsically worthy actions rest in Ross's sense maybe plausibly claimed to have intrinsic worth or value of the same kind that doing our duty as a result of expressing them does; reflective equilibrium, or careful and reflective deliberation, or the intuitive apprehension of moral facts have intrinsic worth in just the same sense as their resultant principles do; rational human nature has the same kind of intrinsic value as the imperatives that express it do. These things have intrinsic value in the sense that we would accord them moral worth even if they were not related to other CPVs as their value-conferring properties, and independent of any valuable consequences they may or may not have. We think it is important for persons to be reflective and rational and for moral relations to obtain, even when the outcome is not one we would have chosen, just as we think it is important to be happy independently of the outcome

doing so may have. This is not to deny that we may need to abdicate any one of these states if the outcomes prove to be disastrous. But we would do so with reluctance, just as we would when forced to give up anything of intrinsic worth. That these properties themselves are intrinsically valuable, or could arguably be so relative to some theory, explains why they confer value on their carriers.

If these value-conferring properties of CPVs are themselves CPVs, there is no difference in metaphysical structure between these properties and any other CPVs. These too may have their value-conferring properties that may be either further intrinsically valuable characteristics, or other CPVs that can be ascribed to them as properties. Thus, for example, the fact that all human beings strive for happiness may confer primitive value or worth on happiness; that friendship and aesthetic experience are sources of happiness may confer primitive value on friendships and aesthetic experience. That fulfilling our obligations rests on morally significant circumstances may confer primitive value on fulfilling our obligations; and that morally significant circumstances reflect rational human nature may confer primitive value on morally significant circumstances; and so on.

Of course these properties always bear a special value-conferring relationship to those CPVs of which they are properties, as stipulated in some particular normative theory. And it is likely that, in general, no such carrier would be a carrier of primitive value without its particular value-conferring properties. Happiness, for example, would not be a CPV if it were not so important to people to attain it. Nevertheless, happiness is no more or less a CPV than the fact of people's aspiring to attain it, as in Hegel's normative theory.[29] For both could occupy the role CPV within some normative theory. Both could confer value or worth on the actions, institutions, or practices that promoted them.

In general, that value-conferring property of a CPV which is itself such a carrier is no more or less of a carrier of primitive value relative to some value theory than that on which it confers value. Since value-conferring properties of CPVs are no more or less diverse in metaphysical structure than any other CPVs, such things as morally significant circumstances, the expression of rational human nature, that all human beings should strive for some one thing or state of affairs, and reflective equilibrium or deliberation can all serve as

[29]Hegel's theory as explicated in *The Philosophy of Right* has often been interpreted as holding as carrier of primitive value not welfare, but the common aspiration to welfare on the part of all members of society (see the essays by Ilting and Plamenatz in *Hegel's Political Philosophy*, ed. Z. A. Pelczynski (New York: Cambridge University Press, 1972). Central to Hegel's conception of the rational Will is the notion that all individuals concur in the adoption of this communality of purpose as itself the highest good; see Hegel, paragraphs 151-55, 257-61; and also my "Property and the Limits of the Self," *Political Theory 8*, 1 (February 1980), 39-64.

intrinsically valuable ends as well as any others, and they can serve equally as the subject of deontological prescription as well as the content of final ends. For example, the expression of our rational human nature is just as plausible as a desired end we may wish to achieve as it is as that which we may view ourselves as directly obligated to do; reaching reflective equilibrium is as likely a candidate for a state we may strive to achieve as it is for a duty we must fulfill as part of action morally. We can express this general truth by saying that those CPVs that are value-conferring properties of other CPVs are indistinguishable in metaphysical structure, or *metaphysically indistinguishable*, from other such carriers. Any constraints on their use or arrangements within some normative theory are a function of their content alone. So final ends (2) in Figure 3 are metaphysically indistinguishable from value-conferring properties (3).

But if it is characteristic of deontological theories that (1)-type elements in Figure 2 occupy position (2), and if (2)-type CPVs would not be such without their value-conferring properties (3), which are similarly CPVs, then relationship (β) in deontological theories is equivalent to relationship (α) in consequentialist ones. For deontologically prescribed actions, institutions, and practices ((1)=(2)) are only provisionally valuable relative to the further CPVs (in position (3)), just as consequentially prescribed actions are, relative to the ends they promote. Thus we can adumbrate the structural equivalence of consequentialist and deontological theories as follows:

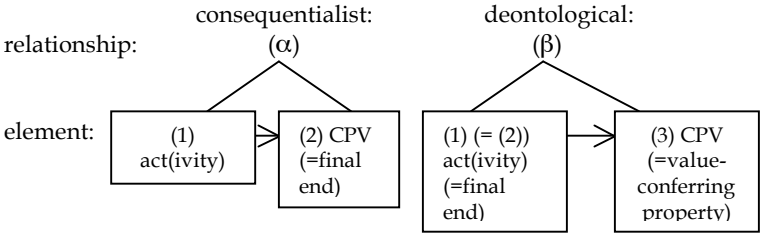


Figure 5. Structural Equivalence of Consequentialist and Deontological Normative Theories

Here we might characterize both relationship (α) in consequentialist theories and relationship (β) in deontological ones as "provisional on the promotion of." I omit independent treatment of (β) in consequentialist theories and (α) in deontological ones, since the arguments of Sections 3.1-3.a conjointly imply their susceptibility to the same line of reasoning.

So, for example, Utilitarianism implies that the commitment to keeping promises is to be abdicated if it does not lead to the greatest amount of happiness possible, whereas Rawls's Social Contract Theory implies that the

two principles of justice are to be abdicated if one would not choose them in a state of reflective equilibrium,[30] and Ross's Intuitionism implies that the list of prima facie duties is to be abdicated if they do not in fact rest on morally significant circumstances (20-28). Just as the moral rightness of some state of affairs depends in a consequentialist theory on its relation to an independent CPV, similarly the moral rightness of some state of affairs depends in a deontological theory on its relation to a similarly independent CPV, namely that value-conferring property of the act, institution, or practice itself.

3.3.2. Provisional Value

To the structural equivalence thesis one may object that even if structural similarity is conceded, structural identity must be denied. For a consequentialist normative theory construction stipulates a CPV as a final end, relative to which the value of morally right states of affairs are not only provisional, but irreducibly instrumental, whereas the morally right states of affairs prescribed by a deontological theory can never be merely instrumental in this way. To answer this complaint we need to scrutinize more closely the terms we used to describe relationship (α) in consequentialist value theories.

First: what does it actually mean to call morally right action *instrumentally* valuable over and above what it means to call it *provisionally* valuable in the sense already explained? Does it mean that the actions, practices, or institutions *promote* or *conduce* to the further, independent CPV in a consequentialist theory but not a deontological one? Surely this is not what it means. Just political institutions, for example, that may be claimed to be intrinsically valuable because they express rational human nature, noncausally promote or conduce to that value which they express, just because they express it. Keeping one's promise, if intrinsically valuable because doing so can be consistently willed as a universal law of nature, noncausally promotes the value of consistently willing the maxims of action as universal laws of nature, just because it exemplifies this value. As we have already seen, "to promote something" need not mean only "to cause it to come into existence." I can promote good music by playing it, or promote the display of affection by displaying it myself, even if neither action has any further causal consequences that are relevant to its promotion. And if I do not cause my action to come into existence, I do not cause that which it promotes to do so either. This is just to repeat that things can be promoted by being expressed, realized, or exemplified, as well as by being caused (cf. (3.3.ii.c). This has nothing to do with instrumental value.

Earlier something described as instrumentally valuable was characterized as a *means* or *instrument through which* its CPV was realized. Fulfilling one's obligations is not, strictly speaking, a means or instrument through which

[30]Rawls, *op. cit.* Note 4, pp. 19-20.

morally significant circumstances are expressed. But what would count as instrumental value strictly speaking? Sidgwick claims that friendship was an important means to the Utilitarian end.[31] But friendship cannot be *strictly speaking* a means or instrument through which happiness is achieved. Only genuine instruments, such as machines that stimulate the pleasure centers of the brain, can be means or instruments in the strict sense. Certainly we are free to view friendship in this way, metaphorically speaking; and in Chapters X.6 through XII I look more closely at some metaethical views that conceive certain key activities similarly. But the same metaphoric liberality then entitles us to view fulfilling our obligations as a means or instrument through which morally significant circumstances are expressed as well. And we need not be consequentialists to do so.

So it appears that so, far, there is reason to suppose a structural equivalence between consequentialist and deontological value theories after all. In both cases, the moral value of action, institutions, and practices have only provisional value relative to their carriers of primitive value – whether the favored normative theory is consequentialist or deontological.

3.3.3. Causation

Earlier, consequentialist theories were represented as insisting upon a causal relation between that which is morally right and the CPV it promotes (3.3.i.c), whereas a deontological theory was supposed to make this relation noncausal and constitutive (3.3.ii). But a consequentialist value theory must accommodate a noncausal constitutive relation between a morally right state of affairs and its independent CPV, and a deontological theory must accommodate a causal relation between a morally right state of affairs and its CPV. If a consequentialist value theory ruled out all such noncausal and constitutive relations, it could not be morally right within a consequentialist theory to promote happiness *through* friendship, or to make someone happy *by*[32] arousing his competitive tendencies at chess, or to promote human perfection *by* developing and exercising one's talents. In each such case, the morally right action is related to the carrier of primitive value as a constitutive part and not as a causal antecedent. But a reasonable consequentialist will rightly exhort its performance nevertheless.

Indeed a consequentialist value theory that consisted only of causal relations would be impossible because it would require us to cause the desired end, but never to participate in it through our own actions or experiences. For example, we might cause happiness to occur, but could do nothing that would be constitutive of being happy. This would imply, first,

[31]Henry Sidgwick, *The Methods of Ethics* (New York, N.Y.: Dover, 1966), p. 437. For a discussion of this claim see Chapter XI.

[32]in Goldman's sense (*op. cit.* Note 28, pp. 5-6, 20-21).

that we would not be permitted to cause ourselves to be happy; second, that no other agent who consciously accepted this theory could permit herself to be caused by anyone else to be happy, since in either case the effect of the action would be that precisely those agents who are only to cause happiness themselves participate in happiness. This is ridiculous. Hence no consequentialist value theory can plausibly rule out constitutive relationships between elements (1) and (2), and this means that (α) must include identity relations, just as do deontological theories.

Similarly, if a deontological value theory ruled out all causal relations between morally right states of affairs and their independent CPVs, it could not prescribe as morally right an action because it effected rather than expressed the theory's CPV. If such a theory contained no causal relations between its primary elements at all, we would be prevented from making any appeal to consequences that were also value-conferring properties of the CPV in order to decide what to do. For example, suppose the obligations to tell the truth and to refrain from harming others were to conflict under certain circumstances. Suppose also that refraining from harming others caused rational human nature – the agent's, the potential victim's, and the potential victim's associate's – to be expressed; whereas telling the truth under these circumstances merely caused human malevolence and spitefulness to be expressed. In deciding what to do, we would be unable to appeal to these consequences even as a tie-breaker. No consideration of the form,

> If fulfilling moral obligations is intrinsically valuable because doing so expresses rational human nature, then to choose between two such conflicting obligations that one which, under the circumstances, causally undermines the expression of rational human nature vitiates the point of fulfilling moral obligations. So I should choose the other one.

would be acceptable. This, too, is ridiculous. A normative theory that rules out this kind of reasoning is not one that any deontologist – no matter how pure – would be likely to adopt.[33] Hence no plausible deontological value theory can rule out causal relationships between elements (2) and (3). This makes the relationship (β) comparable to relationship (α) in consequentialist theories.

Thus the consequentialist can no more claim a value-theoretic monopoly on causal relations between morally right actions and their CPVs than the deontological can on noncausal, constitutive relations between them. The particular character of the relation (α) is not determined by whether a theory is consequentialist or deontological in form, but once again only by the content of that theory. Any such value theory must contain both kinds of

[33]Cf., for example, Ross, *The Right and the Good, op. cit.* Note 17, p. 31.

relation in order to be normatively viable.[34] These considerations taken together suggest that if structural equivalence between consequentialist and deontological value theories is in fact lacking, some further, nonshared property needs to be adduced to demonstrate this. And of course it must also be demonstrated that this property is not itself particular to the content of some one such theory.

4. Two Metaethical Attitudes

All along, the focus has been on the structure and content of normative theories, independent of the metaethical attitudes and pronouncements ethicists make about those theories. If my treatment of Anscombe's thesis has been correct so far, the basis for the consequentialist/ deontological distinction is not to be found in any property of normative theories themselves, but rather in those metaethical attitudes expressed by Anscombeans. So I now want to consider those attitudes. I show that they are based on mistaken beliefs about the applicability of this distinction to normative theory, and on psychological attitudes that would be better expressed in a very different distinction.

Anscombeans often seek support in the self-evident fact that there is, after all, a disagreement between someone who thinks it is always wrong to

[34]Some think the telling difference between consequentialist and deontological value theories consists in the status they accord to moral injunctions, whether causal or constitutive. They think a consequentialist theory treats them as disposable rules of thumb, whereas a deontological theory regards them as universally binding laws. Peter Railton expressed this view. But it applies only to dummy consequentialist and deontological theories respectively, not to any real ones; and even then only to their practical, not their value-theoretic parts. We have already seen that the value-theoretic part of a normative theory supplies no action-guiding directives on how we should promote or realize that which has moral worth, much less on how often we should do so. On the other hand, the practical part of any viable consequentialist theory must recognize that certain actions are in fact always morally obligatory, not only because in fact they might always best promote the value-theoretic good; but also because they are most reliable in cases where we cannot know what act would do so – which, as we have already seen, is itself a permanent feature of practical consequentialist injunctions. So practical consequentialist prescriptions are frequently universal in character (Sidgwick and Moore are particularly explicit about this). Similarly, practically viable deontological prescriptions recognize that value-theoretically prescribed duties cannot always be successfully fulfilled. As we have already seen, they may conflict or they may fail to be completed successfully. In these cases a practicing deontologist is prepared to perform that action which on the whole best conforms to the theory's value-theoretic prescriptions, and also to revise her conduct in case it turns out not to serve this purposes. So practical deontological prescriptions frequently have the character of rules of thumb (cf. Ross, *The Right and the Good*, pp. 30-32). That both consequentialist and deontological practical prescriptions must include both universal laws and rules of thumb follows directly from the prescriptive indeterminacy thesis.

lie, regardless of the causal consequences of doing so, and someone who thinks it is only wrong to lie when lying does not promote general welfare. Similarly, those who care more about conforming their behavior to clear-cut moral prescriptions that about making themselves and other people happy are clearly at odds with those whose priorities are the reverse. From these facts it is often concluded that there is a genuine disagreement between consequentialists and deontologists after all.

But this conclusion does not follow. That is, it does not follow from the fact that people have different moral priorities, or accord greater or lesser moral value to different states of affairs, that they must make a commitment to consequentialism or deontologism. For as we have seen, any such content may figure in consequentialist or deontological theories indifferently; and their respective structural relationships are equally unhelpful in classifying one's moral convictions in one way rather than the other. So it will not do to argue here that it is just my preferring never to lie rather than effecting the general welfare that *makes* me a deontologist. For my adoption of the value of never lying is as such neutral between adopted ends and the means to their achievement, and neutral between carriers of primitive value and those states of affairs that promote them.

So our disagreements about the relative importance of performing different actions or achieving different ends shed no light on the consequentialist/ deontological distinction. All it proves is that people do indeed differ about whether it is more important to tell the truth than to be happy, to distribute goods and service justly than to satisfy desires, and so on. But this fact is uncontroversial. All these possibilities are metaphysically indistinguishable values to which different individuals may assign different weights without thereby providing evidence for their consequentialist or deontological proclivities. This is not to argue that people do not have such proclivities. Psychologically and professionally, a great deal may turn on whether one fancies oneself to be a consequentialist (tough, hard-nosed, practical but idealistic) or a deontologist (stern, uncompromising, virtuous but not intolerant). It is just to claim that such self-conceptions find support in neither the values nor the structure of the normative theory any such individual is likely to hold.

Of course part of the intensional attitudes of some such ethicists include not only these values, but in addition the conviction that some particular value is an end to be achieved, or a means to some such end, or descriptive of an intrinsically valuable action irrespective of the ends it may promote. Such an individual may maintain an explicit lack of interest in, say, the consequences that particular prescribed actions may promote, or, alternatively, in the particular means undertaken to achieve some desired end, and identify herself respectively as a deontologist or consequentialist on these grounds alone. But if the analysis offered in Sections 1 through 3 of this

chapter are correct, those intensional attitudes toward the components of moral action are simply confused.

However, Anscombeans may then cite the very clear differences in ethical sensibility that often motivate adherence to consequentialism or deontologism.[35] Self-styled deontologists often regard their own imperfect attempts to do what they believe to be right as challenge enough, without incorporating any vision of what would be good for other people into their moral program. They may believe that their primary task is to attend to their own moral behavior, while relying on the essential humanity and rationality of other people as sufficient evidence that they will do the same. This conviction may be explained by the assumption that these two characteristics, of rationality and humanity, are sufficient conditions for inclusion in a general moral community whose continued existence is dependent on the capacity for moral autonomy, i.e. for generating and regulating one's actions in accordance with universal moral laws. Those who exercise this capacity for immoral purposes are then viewed as fully responsible agents to be condemned or punished, but never remade or reprogrammed in ways that would be thought to violate their essential personhood. Deontologists may thus regard as both arrogant and manipulative the consequentialist's eagerness to assume responsibility, not only for his own behavior, but for events and states of affairs that may be only remotely causally contingent on it; and to take on the project of the moral reform of others on a grand scale as part of one's personal moral program.

Self-styled consequentialists, on the other hand, often believe that a healthy sense of sympathy and compassion for other people profoundly demands a commitment to their welfare that may even outstrip one's commitment to one's own. This sentiment may be justified by a broader conception of the moral community that includes all sentient beings, or perhaps all beings with complex central nervous systems. Thus they may be less inclined to differentiate between moral agents based on degree of competence or rationality. They may therefore find unthinkable a morality that requires them to ignore the fact that all moral agents and their behavior are mutually interdependent within a common sociopolitical and causal network, just as all beings and events are within the larger common physical network. They may view as selfish and irresponsible the deontologist's preoccupation with her own moral probity, and willingness to sacrifice the well-being of other people on the altar of moral law.

These are serious attitudinal differences indeed. But they bear no relation to the substance of anyone's ethical views. We have already seen that disagreements over actual normative priorities do not force the commitment

[35]Stephen White's insights and critical comments have helped this and the following paragraph.

to consequentialism or deontologism. A perfectionist defines the moral community in much the same way as the deontologist supposedly does; deontologists often extend the scope of their moral concern just as broadly as the consequentialist (as, for example, in Nozick's theory of animal rights). With the possible exception of those ethicists who hold and act on just that false belief which I am attacking, i.e. that adopting some substantive normative value or priority implies a consequentialist or deontological commitment, there is no evidence to support any correlation between these.

Now the deontologist ascribes moral arrogance and manipulativeness to the consequentialist because of the latter's assumption of moral responsibility for events over which, it seems, only an omnipotent being could have control; and also because of his concern with effecting the welfare of other people, independently of their *prima facie* wishes or collaboration. In Chapter XII I develop in depth this criticism of Classical Utilitarianism specifically. But the deontologist's own aspiration to perfect adherence to the moral law, and apparent disregard for inherent human imperfection and irrationality, may just as easily provide fuel for the accusation of moral arrogance, as may the conviction that the preferred set of moral principles are innately superior to any that are either incompatible with them, not a product of Western culture, or both.[36] Deontological manipulativeness may be similarly demonstrated in the insistence on systematic moral education in case one is not inclined to adopt the favored principles. Here the reasoning may be that one merely needs to, for example, develop one's capacity for moral intuition, achieve a higher level of rational or moral development, or be taught to respect the moral law, in order to estimate these principles at their proper worth.

On the other hand, the consequentialist criticized as selfish and irresponsible the deontologist's concern with personal moral virtue at the expense of general human welfare. But the consequentialist's own selfishness might be just as easily evinced by his insensitivity to the very real desire of other people to determine freely and without outside interference the course of their own lives, and to pursue their own conceptions of the good. Similarly, the consequentialist might demonstrate moral irresponsibility in his willingness to discount or sacrifice the claim of an innocent life if doing so will further social welfare.

The general point is clear. Moral arrogance, manipulativeness, selfishness, irresponsibility, and indeed a host of other vices one might have occasion to ascribe to particular ethicists are not the exclusive preserve of any one type of normative theory, any more than is the moral humility, respect for others, altruism, or sense of responsibility by which the accusers would – and could – presumably characterize their own normative views. These qualities

[36]Some evidence of this conviction can be gleaned from passages in Kant, Ross, Kohlberg, and Rawls.

describe attitudes and psychological dispositions that individuals may or may not have. And these attitudes and dispositions may or may not infect the expression of one's moral convictions. But these convictions themselves are logically independent of both the personality problems and character traits of the individuals who hold them, and of the consequentialist/ deontological taxonomy. So neither these convictions nor the personal traits that supposedly accompany them reflect the supposed difference in moral sensibility that Anscombeans claim.

Now there are certain criticisms of deontological and consequentialist theories often made by members of the opposing camp that have a common ring to them. Consequentialists often claim that deontological theories are guilty of "rule worship," and are essentially unconcerned with people; for they inflexibly prescribe certain actions without regard to how others are affected. They fail to recognize the importance of human well-being as an intrinsic value.[37] Deontologists then typically retort that it is the consequentialist who exhibits an essential lack of concern for people. For consistent consequentialist theories require the sacrifice of the innocent for the sake of some "greater good,"[38] subordinate human rationality and autonomy to the pursuit of this good,[39] and fail to respect personal integrity.[40]

Note that, as usual, the criticisms could be reversed. One could just as easily fault consequentialist theories for paying insufficient attention to human welfare on the grounds that they subordinate individual well-being to the general welfare. One might then go on to argue that a theory that places individual welfare in jeopardy threatens and thereby diminishes the welfare of each individual in the community, hence diminishes general welfare. One could similarly criticize deontological theories on the grounds that a thoroughgoing commitment to general principles of moral obligation undermines the opportunity to exercise individual rationality and autonomy in decision-making on particular occasions, since individual inclinations are in each case subordinated to the principle of conformity to these general normative prescriptions.[41]

---

[37]J. J. C. Smart makes this objection in "An Outline of a System of Utilitarian Ethics," in J. J. C. Smart and Bernard Williams, *Utilitarianism: For and Against* (Cambridge: Cambridge University Press, 1975), pp. 5-6, 72. Also see Jonathan Bennett, "Whatever the Consequences," *Analysis 26* (1966), pp. 83-102.

[38]H. M. McCloskey, "A Note on Utilitarian Punishment," *Mind 72* (1963), p. 599.

[39]Thomas Nagel, "Subjective and Objective," in *Mortal Questions* (Cambridge: Cambridge University Press, 1979).

[40]Bernard Williams, "A Critique of Utilitarianism," in Smart and Williams, op. cit Note 37; also see Rawls, *op. cit.* Note 4, Sections 5 and 30.

[41]W. D. Falk makes essentially this criticism in "Morality, Self, and Others," in Judith J. Thomson and Gerald Dworkin, Eds., *Ethics* (New York: Harper and Row, 1968);

That the objects of these criticisms can be interchanged so easily suggests that it is in fact not the consequentialist or deontological structure of these theories that is under attack, but something else. These criticisms have in common the reproach that the theory under fire is what we might call insufficiently *person-regarding*, i.e. that it ignores or devalues the importance of certain human needs and requirements that are centrally important from the point of view of normative theory: that we should be happy and not miserable, that we should be permitted and encouraged to determine the course of our lives, that the value of different conceptions of individual welfare should be recognized and respected, and that we should be able to be both rationally self-directing and also fully committed to the plans and projects to which we attach value. Unlike the epithets consequentialists and deontologists usually hurl at one another, this reproach is a serious one, for it touches on the most basic rationale for adhering to or constructing a specifically normative theory in the first place. If fulfillment of these needs and requirements is of central significance for human beings, and if the whole point of a normative moral theory is to regulate relations among human beings in a rational and practically effective way, then a theory that is insufficiently person-regarding in this sense can claim very little title to support at all.

Avowed Anscombeans often acknowledge that this criticism presents genuine difficulties for their respective metaethical allegiances. Some consequentialists may respond by incorporating the values of rationality, autonomy, integrity, or respect for persons into the characterization of human welfare as the carrier of primitive value, or as empirically necessary means to the realization of this end. They then worry about how to square the importance of such values with the consequentialist structure of their favored theories. Some deontologists may respond by insisting that as a matter of empirical fact, adherence to moral principles of action conduce to human welfare, while attempting to defuse the suspicion that they have thereby sullied the deontological purity of their theories with a consequentialist justification.

Not all Anscombeans have this response. Some consequentialists accept the charges of scapegoatism, paternalism, or alienation with a shrug, claiming these unfortunate flaws to be the necessary price of practicability. Similarly, some deontologists accept the charges of rule-worship or lack of human sympathy as the necessary concomitants of consistency.

Thus this disparity of response to the criticism does not parallel, but rather cuts across the consequentialist-deontological distinction. On the one side, we find those who attempt to restructure their normative theory so as to

---

reprinted in Hector-Neri Castaneda and George Nakhnikian, Eds. *Morality and the Language of Conduct* (Detroit: Wayne State University press, 1963).

fully accommodate the missing values. The resulting "mixed" views, like the four traditional ones discussed in this chapter, are comparable in emphasizing an essentially person-regarding orientation at the expense of easy taxonomic classification. On the other side, we find those who believe that an essentially nonperson-regarding, or not fully person-regarding orientation is a small price to pay for structural clarity and methodological rigor. We might describe such views as *theory-regarding*, meaning by this that their proponents are prepared to accept without further argument the devaluation of certain of the above-listed needs and requirements – the satisfaction of desire, for example; or personal integrity – because of a deeper commitment to what they perceive as the distinctive structure and method of their theory.

Now the question whether the satisfaction of desire is more or less important than personal integrity, or whether autonomy is more or less important than happiness, is a normative issue to which I attempt no answer in this project. I am not sure it can be answered. But if Anscombe's thesis is as superfluous as I have tried to show, those theory-regarding views that opt for this brand of theoretical purity at the expense of any of these centrally person-regarding values are defending a dummy theory, in more ways than one.

5. "Consequentialism" and the Humean Conception of the Self

So far we have seen that Anscombe's consequentialist/ deontological distinction is too superficial to capture the richness and conceptual complexities of actual normative theories. When we examine such theories in detail, the abstract generalizations Anscombeans tend to make about them evaporate on contact. But this is not to claim that Anscombe's thesis has no significant application at all. On the contrary: it points to a deeper distinction in metaethical content that normative theories presuppose. Specifically, as I now argue, Anscombe's use of the term "consequentialist" in fact denotes normative theories that presuppose the Humean conception of the self. For so-called "consequentialist" normative theories presuppose a motivational model based on desire rather than one based on intention and will, and a structural model correspondingly based on the maximization of utility as the basic criterion of rationality.

Consider first the desire basis of such theories, beginning with our conception of action. Intention-descriptions are semantically equivalent to act-descriptions: they differ in their referents, but not in their meaning. So although we know the metaphysical difference between an intention (a goal-directed mental – i.e. intensional – state) and an action (an extensional physical event consisting in overt behavior), the contextually isolated statement that denotes one in one context can equally well denote the other in a different one. This means that when we are describing an action, we are thereby describing the agent's intention in performing the action: to take a walk, for example; or to feed the cat. Since the description of an action

actually performed necessarily includes a description of its most immediately intended effects – i.e. the intensional goal of the action, a motivational model based on intention and will in some cases equates the two: the action actually performed will be the same as the immediate effects its agent intended.

A Kantian, so-called "deontological" normative theory assigns moral value to certain act-types in virtue of their immediately intended effects: telling the truth, for example; or helping others – even though the immediate actual or more remote actual effects of such act-tokens may backfire. The description of these intended effects fixes both the goal of the action, and thereby the identity of the action actually performed, or act-token. A Kantian normative theory thus depends on the semantic equivalence of act-descriptions and intention-descriptions in identifying its set of morally valuable acts. It stipulates certain kinds of intentions as the primary object of moral worth, irrespective of their actual immediate or remote effects. So, for example, if I intend to tell the truth but garble my actual utterance so badly that I am unintelligible to those around me, I have still performed the right action because of the intention behind my behavior. On this view, an action actually performed is morally valuable if and only if the effects it immediately intends are.

By contrast, a "consequentialist" normative theory purports to require a sharp distinction between an action's intended effects and its actual effects in all cases, because it must retroactively evaluate the moral worth of all actions with reference to the value of the effects they have actually achieved, regardless of their intention. Of course it accepts the conceptual equivalence of act-description and intention-description, since this equivalence is a conceptual truth; and so the conceptual identification of the physical action performed with its immediately intended effects. If there is a piece of lettuce caught in your beard, then my utterance,

There gibt une stuck kopfsalat stuck in your barbe

can be described as my telling the truth. But in a "consequentialist" normative theory, an action's immediately intended effects are neither sufficient nor necessary for determining its moral value. Instead, such a theory stipulates as a source of moral worth something that stands outside this equivalence, namely the actual effects the agent's physical behavior has caused. The question is not whether my garbled warning about the lettuce in your beard counts as telling the truth, but rather whether my telling this truth has beneficial or harmful consequences. If these consequences are beneficial according to the values of the normative theory in question, then the action is worthwhile, irrespective of the agent's intention in performing it. If they are not beneficial according to the values of that normative theory, then the agent's good intentions cannot make them so.

An intention- or will-based model of motivation does not satisfy this requirement. But a desire-based model does, because it observes in all cases the distinction between intended effects and actual effects, and utilizes this distinction in a metaethical criterion for ascribing value to action that is independent of the agent's intention. This criterion begins with desire. I can desire the effects of an action without intending them, and I can intend the effects of an action without desiring them. An example of desiring the effects of an action without intending them would be desiring so desperately to stop smoking that I would welcome a neurological implant or stroke that might effect this; but lack the resolve to do anything that might have this effect myself. Here I desire certain actual immediate and more remote effects of an action or action-plan, but because I have no intention of carrying it out, I intend no effects of either kind. An example of intending the effects of an action without desiring them would be deliberately greeting my enemies in the halls every day, even though it makes me physically nauseous to do so and I know it will have no beneficial consequences (they will hate me more, not less, for my apparent equanimity, and I will have to contend with chronic gastric disturbance). In this case I intend some of the immediate effects of the action but desire none of its actual effects, whether immediate or remote.

The former case is one in which I do desire certain actual immediate or remote effects of action, whereas the latter case is one in which I do not desire them. By contrast, in the former case I do not intend the effects I desire, whereas in the latter case I do intend effects I do not desire. So in both cases, the intensional objects of my desire are conceptually independent of the intensional objects – or effects – I intend by acting. Indeed in these particular cases, the objects of desire and the objects of intention are at odds. But in all cases, desire provides the first half of a conceptually contingent criterion for assigning value to the actual immediate and remote effects of an action that is independent of the agent's intention in performing it.

The remaining half of that criterion of value is the notion of satisfaction. The criterion of desire-satisfaction – *actual* desire-satisfaction, regardless of intention or resolve – furnishes the standard against which the actual consequences of action are evaluated by the "consequentialist." Of course the kind and content of desire will vary with the "consequentialist" theory in question: The Classical Utilitarian evaluates uncorrected desire, whereas the Millian Utilitarian evaluates educated desire, the Brandtian Utilitarian therapeutically informed desire, and the Marxist ideologically enlightened desire. The Act-Utilitarian desires individual happiness- or pleasure-events, whereas the Rule-Utilitarian desires happiness- or pleasure-producing social rules, the Marxist the classless society, and the Perfectionist the full flowering of human capacities.

It is because actual desire-satisfaction can be distinguished and detached from the agent's intention in acting – regardless of her motives – that it can

provide a criterion for evaluating act-consequences that appears agent-independent. And it is the seeming agent-independence of this criterion that makes plausible the conviction that any such actual consequence – happiness, pleasure, the classless society, human perfectibility – has value independent of what any human being might think about it. It is this conviction, in turn, that rationalizes the manipulative social strategies I dissect in Chapter XII below. I do not here take a position on whether or not intrinsic values exist; they may well exist. My claim is a more modest one: That an actual consequence of action satisfies an agent's actual, uncorrected, educated, therapeutically informed, or ideologically enlightened desire, independent of that agent's intention, does not by itself suffice to demonstrate the independent value of that consequence. That the criterion of desire-satisfaction evaluates act-consequences independent of agent intention does not lend the resulting values agent-independent authority.

It is because a desire-based motivational model requires us to distinguish between the intention behind an action and the value of its actual effects, whereas an intention- or will-based model does not, that the Humean belief-desire model of motivation implicitly stands behind all such normative "consequentialist" views. Desire-satisfaction provides a conceptually contingent criterion for evaluating the moral worth of an action that is independent of the agent's intentions in performing it, and thereby lends it an aura of independent value *simpliciter*. I pursue further the contrast between the intention-based analysis of action that grounds the Kantian tradition in action theory and the desire-based analysis that grounds the Humean tradition in Chapter IX.3.4 below.

In theory there are, of course, other conceptually contingent candidates for the metaethical criterion of value-ascriptions to action besides desire-satisfaction – for example, what God commands, what is human, natural, rational, or divine. But for so-called "consequentialists," these alternatives will not do unless they provide a necessary motivational correlate; and desire and intention are the only two plausible candidates for motivation. Without an equivalence between what is valuable and what motivates action, so-called "consequentialists" cannot make hortatory appeal to the valued consequences of action they prescribe as reasons for performing it. Of course such "consequentialists" might believe there were such reasons, even if they recognized that these reasons had no motivational appeal to human agents. They then would have, not only no viable normative theory, but no viable reasons for action in their metaethical arsenal. Given the already shaky metaethical standing of "consequentialism," it needs internalism to enhance its credibility.

This means that the metaethical criterion for assigning normative value to outcomes – whether that assigned value is happiness, pleasure, human perfection, or a classless society – has to be the same as the criterion for

assigning causal efficacy to human motivation – whether that motivation is passion, desire, emotion, or sentiment – to pursue those outcomes. Among human passional states, only desire (and those states reducible to it) can satisfy this dual role, of functioning both as a conceptually contingent source of value and a conceptually necessary source of motivation. For unlike excitement, anger, happiness, fear, joy, shame, or other passional states, desire is the only one that necessarily and always carries an intentional object. Only desire, among such states, can always be relied upon to provide intentional direction to human behavior. The desire-state itself is stipulated to confer value on the intentional object, and the valued intentional object in turn is stipulated to motivate the action believed most efficiently to realize it. This just is the Humean belief-desire model of motivation already examined in Chapter II.

The inference to the necessity of a utility-maximizing model of rationality of the kind discussed in Chapters III and IV for so-called "consequentialism" is straightforward. If obtaining more valued consequences of action and minimizing the expenditure of resources in their service were not better, this would imply some further constraint relative to which the value of those consequences themselves were restricted. This further constraint itself would have to be, *ex hypothesi*, nonconsequentialist in nature. This, in turn, would contradict the "consequentialist" first principle that actual consequences alone determine the moral value of actions. So "consequentialism" presupposes both the belief-desire model of motivation and the utility-maximizing model of rationality. It thus presupposes the Humean conception of the self. *Contra* Anscombe, it is this view that is the anachronistic one; and in what follows I argue that those who seek to found their moral theories on its basis effectively shoot themselves in the foot.

Chapter VI. The Problem of Moral Motivation

I have just argued that Anscombe's consequentialist/ deontological distinction is without substance, and further that the actual metaethical foundation of so-called "consequentialist" normative theories is in fact the Humean conception of the self. The remainder of this volume of the project therefore scrutinizes in greater detail those leading moral theories that presuppose this conception. In Chapter I, I also claimed that arguments defending the centrality of rationality in the structure of the self presuppose the value of rationality as the self's defining element; and that thus valuing rationality entails one's readiness, first, to recognize it as definitive of the self; and second, to valorize its character dispositions. The Humean conception of the self takes an analogous valuational stance toward desire. It views desire as the defining element in the self and valorizes its character dispositions (to feel satisfaction, frustration, satiation, discontent, pleasure, pain, etc.) accordingly. But I argued in Chapters II through IV that to be defined and moved by desire alone was to be defined and moved by essentially egocentric considerations. This chapter examines desire, its centrality in the structure and motivation of the self on the Humean conception, and the first of three central and *in theory insurmountable* problems in metaethics to which the Humean belief-desire model of motivation gives rise. The first of these three problems is that of moral motivation.

Section 1 formulates the problem, namely, that on the Humean conception it seems impossible to be moved to act in others' interests on the basis of moral considerations alone. By distinguishing between personal and impersonal desire on the one hand and between self-interest and self-direction on the other, it then refutes Bernard Williams' and Annette Baier's claim that we can be motivated by desire without being motivated by self-interest. Section 2 examines Rawls's distinction between object-dependent, principle-dependent, and conception-dependent desires; and rejects his more fundamental distinction between motivational force and psychological strength on which the account of principle-dependent desires is based. It concludes that this complex set of distinctions neither solves nor escapes the problem of moral motivation that besets all Humean accounts. Section 3 recurs to the distinction between personal desire-satisfaction, which is inherent in the belief-desire model of motivation, and pleasure, which is only contingently related to it; and rejects the thesis that personal desire-satisfaction entails that other-directed desires are reducible to self-directed ones. Section 4 examines other-directed desires that are selfish or self-indulgent without being self-directed. Finally, Section 5 considers what it would mean to be motivated by other-directed considerations that are independent of desire-satisfaction. It introduces the example of the unprotected whistleblower as a paradigm case of this sort of genuinely

disinterested type of motivation, and rebuts the Humean attempt to reduce such an example to a rather less inspiring case of concealed egoism. It thus prepares the way for a more in-depth treatment of the whistleblower in Volume II, and a more satisfactory explanation of this phenomenon in transpersonally rational terms.

1. Self-Interest and Other-Direction

The belief-desire model of motivation generates the problem that moral motivation in any meaningful sense does not seem to be possible within the designated constraints of this model. "Moral motivation" usually means non-egocentric motivation: motivation independent of self-interested or personally opportunistic considerations such as comfort, convenience, profit, or gratification. Non-egocentric motivation is by moral considerations alone – transpersonally rational appeals to intellect and conscience – that inspire us to act in others' interests even when this requires sacrificing or ignoring our own. Since the Humean conception stipulates desire as the sole conative impetus to action, it is on the face of it hard to see how or where such moral considerations might effectively function. It would seem that self-interested motivation is the only kind the belief-desire model recognizes.

A motive is *self-interested* if it includes an interest the self takes in its own condition. Some Humeans, such as Bernard Williams and Annette Baier, argue that one can be motivated by desire without being motivated by self-interest.[1] Since desire can take a variety of objects, including other-directed ones, one can be motivated by altruistic desires such as benevolence or compassion that are not self-interested at all. Or so the reasoning goes (in Volume II, Chapter VI.4 I offer an analysis of compassion that disputes this.).

But self-interest, in turn, can be dissected into short-term personal gain (i.e. immediate self-interest) and long-term personal gain (i.e. prudence). Satisfying a desire is one kind of personal gain, and the frustration of a desire is one kind of personal loss. To anticipate the satisfaction or frustration respectively of personal desire is one kind of anticipation of a short- or long-term personal gain or loss respectively, and we have already seen in Chapter II that without such anticipation we cannot be said to desire the object or state of affairs in question. Since such anticipation is, in turn, one kind of interest the self may take in its own condition, the satisfaction of personal desire is one kind of self-interested motive. So the Humean conception of the self in effect asserts that only self-interest can motivate us to act to promote others' interests. It therefore "solves" the problem of moral motivation by in effect denying that genuinely moral motivation is possible.

---

[1]Williams' and Baier's views are examined in greater depth below, in Chapters VIII and XIII respectively.

1.1. Personal vs. Impersonal Desire

A Humean may attempt to evade this conclusion by claiming that other-directed desires such as benevolence or (on one analysis) compassion are impersonal rather than personal desires. An *impersonal desire* would be one whose object is desired independently of any personal relation it or its realization might or might not bear to the agent who desires it. However, even to state the definition reveals its incoherence. If the agent does indeed desire that object, then it necessarily bears a personal relation to the agent as that which satisfies that agent's desire. Of course the agent might not know that the object of his desire has been satisfied; or might not be instrumental in realizing the object of his desire. But that the agent bears no epistemic or causal relation to the object of his desire does not imply that he bears no such personal relation. If I desire that American white supremacist youth groups see the error of their ways, then if they do, my desire is satisfied even if I have not yet learned of it. And if I have, then I will be personally satisfied even if I had nothing to do with bringing this desired consequence about.

Thus any desire an agent has is a personal one, including altruistic desire. Any such desire is personal rather than impersonal because it compels the subject's attention to her own state of personal insufficiency (or, literally, want) in relation to an envisioned object she desires; that is part of what motivates her to ameliorate this insufficiency by acting to satisfy the desire. This is true whether the object of desire is a jelly doughnut, the alleviation of another's pain, or ascertaining once and for all the age of the universe. In all of these cases, the desire draws one's attention to something that is missing (or wanting) in one's present state: the taste of a sweet, sticky pastry, or the awareness of another's restored comfort, or the knowledge of an important fact, respectively.

So in all such cases, the experience of desiring thereby grounds an agent's point of view in a relation between his personal awareness of his present condition at a certain time and place at which the desire occurs, and an envisioned state of affairs that he locates at a future time and place at which his desire is satisfied. Desires and their objects situate us as agents in a comprehensive space-time matrix at two or more points which we traverse through the actions we take in order to realize them. Desires always embed us – and sometimes trap us – in the personal point of view. That is why, regardless of their content, they can cloud or bias our attempts at impersonal, impartial, or objective – i.e. transpersonal – judgment. I described this condition in Chapter II as "funnel vision."

1.2. Other-Direction

Humeans might concede that all desires are personal, yet deny that satisfying other-directed desire is a species of self-interested motivation. They might say that if it is the other-directed *object of desire itself*, and not the

satisfaction one experiences by realizing it, that one is motivated to obtain, there is then nothing self-directed about it. But first, drawing too sharp a distinction between the object of desire and the satisfaction of that desire can lead to conceptual confusion. After all, an object of desire is not an object of disinterested contemplation that somehow draws one toward its realization as though it were a magnet, such that the agent has no conative personal investment in that realization. If it is an object of *desire*, then one does have a personal conative investment in its realization, namely that one's desire for it be satisfied. Certainly this is consistent with an agent's intensional focus on the realization of that object rather than on the personal experience of satisfaction that results. Second, my claim that the satisfaction of personal desire is a species of self-interest in any case does not thereby equate personal desire with self-directed desire. Whether a desire is self-directed or other-directed depends on the intensional content of the object of desire.

A desire is *self-directed* if its object represents as desirable some aspect of one's self. It is *other-directed* if its object represents as desirable some aspect of something other than one's self. Of course an object of desire may include representation of both aspects. But I confine investigation to the simpler, unipartite case. Both self-directed and other-directed desires may be either benevolent or malevolent. An example of a benevolent self-directed desire would be the desire to prosper. An example of a malevolent self-directed desire would be the desire to degrade oneself. So a malevolent self-directed desire would still be self-interested because it would include an interest in the self's own condition. A desire can be both self-interested and self-destructive, since the interest the self takes in its own condition need not be healthy.

Similarly, other-directed desires may be benevolent, as when one desires that one's loved ones prosper; or malevolent, as when one desires that one's enemies be crushed; or neutral, as when one desires that the age of the universe be settled once and for all. What links all of them, both self- and other-directed, benevolent and malevolent, as motivationally effective is the anticipation of the desired object as a source of personal satisfaction – i.e. of wholeness and sufficiency restored; and its absence as a source of personal frustration. It is the desire that this object be realized that moves one to action in its service (respectively, for example, to invest in mutual funds for oneself, or for one's loved ones; or sell worthless junk bonds to one's enemies; or write an irate letter to *Scientific American*).

That the object of an other-directed desire is envisioned as a source of personal satisfaction does not imply that one desires the experience of satisfaction rather than the particular object of desire itself, or more than that object, or as much as that object, or even anywhere near as much as that object. Just to make the point makes the distinction as clearly as we need to make it.

On the other hand, merely to envision something as a consequence is not the same as desiring it, even if one envisions it as in some way satisfying to oneself. I can envision something as satisfying without being moved to achieve that satisfaction (right now, for example, I am envisioning the satisfaction of floating in the shallow water of the beach at Negril, but actually I am completely satisfied sitting here at my computer). But I cannot be said to desire that thing if I am not so moved.

However, that the object of an other-directed desire is envisioned as a source of personal satisfaction does imply that if no such satisfaction were anticipated, one also could not be said to desire that thing. Even if one assigns no value to the satisfaction itself (and of course one can be satisfied by something, such as a chocolate-covered cherry, without valuing it), it is still conceptually impossible to desire that object unless one anticipates the satisfaction as a concomitant of it. So I can envision a satisfaction without desiring it, but I cannot desire a satisfaction without envisioning it as such. Envisioning an object or state of affairs as personally satisfying is a necessary but not a sufficient condition of desiring it. This is true by definition of "desire," and holds whether the desire is self-directed or other-directed, benevolent or malevolent.

1.3. Interest in vs. of a Self

In this sense the satisfaction of all such desires do fall under the rubric of personal gain (similarly, one can gain something, such as rental property, without valuing it), and so more generally under the rubric of self-interested motivation (therefore, it can be in one's interest to gain something, such as a painful lesson about human nature, without valuing it). According to this classification, to realize an other-directed desire is, at the very least, a short-term personal gain because of the personal satisfaction this will bring, even if it also requires personal sacrifice. So, for example, a person who desires to devote her life to the liberation of her country receives personal satisfaction from doing so, even if she loses her life in the effort. And because she anticipates receiving this personal gain, her devotion is self-interested. One can be self-interestedly motivated to satisfy an other-directed desire that demands the sacrifice of one's life because survival is not necessarily the only, or the most important, or the highest interest in the condition of one's self one can take. Interests in the condition of one's self that may outweigh one's interest in its survival may include interests in its integrity, rectitude, or surrender. Indeed, even one's interest in an anticipated personal gain may conflict with one's interest in survival, as when one endangers one's life for the sake of a high-risk but high-paying job. Similarly, the self may take an interest in satisfying its other-directed desires which is just as strong as, or stronger than, its interest in satisfying its self-directed desires. The sense in which the Humean model of motivation implies self-interest as the most

plausible explanation of why we ever act to promote another's interests does not thereby confine that explanation to a single, monolithic motive.

The interest a self takes in the satisfaction of its other-directed desires is a genuine interest *in* the self, and so is distinct from what John Rawls calls interests *of* a self.[2] These are discussed at greater length in Chapter X. But briefly, interests of a self include other-directed interests and beliefs that do not necessarily involve desires at all. Particular moral or religious or political convictions might be among the other-directed interests of a self that bear no necessary relation to the self's interest *in* satisfying its other-directed desires. So, for example, a belief in the welfare state, or in retributive justice might be among the interests *of* one's self, yet bear no necessary connection to one's interest *in* satisfying one's other-directed desire that one's friends prosper.

However, interests of the self that are also interests in the self need not be interests in satisfying the self's desires. They can also be interests in adhering to the self's abstract convictions. Even though such beliefs are not specifically self-directed in content, they can be of interest to the self because they are its own. In such a case the self takes an interest in them because it authors and owns them. Since authorship and ownership of a belief is a condition of the self, the self is taking an interest in its own condition even when the content of the conviction is not specifically self-directed. Therefore one can have a self-interested motive in adhering to abstract or other-directed conviction.

For example, Michael Walzer distinguishes between a would-be leader of the oppressed whose actions are justified by his ideology, and one whose actions are justified by the acceptance of her ideology by the oppressed as a set of terms in which their interests are adequately expressed.[3] The first, he points out, is obligated only to himself and those who share his commitment, whereas the second is obligated to the oppressed group from whom she seeks ideological legitimacy. Both leaders are motivated by interests of a self and neither is motivated by desire. But the first leader takes an interest in the condition of his self, namely that *its* ideology furnish the justification of his actions. The second, by contrast, takes a greater interest in the condition of other selves, namely that the oppressed accept her ideology as adequately expressing *their* interests. The first is inspired to lead by his interest in his ideology; the second by her interest in the legitimacy of her ideology among those she leads. The first is motivated to action by an interest that is of and in the self, whereas the second is motivated by an interest of the self that is other-directed. Walzer's distinction shows that there is nothing inherently sacred either about interests of a self that are not desires, or about beliefs that

---

[2] John Rawls, *A Theory of Justice* (Cambridge, Mass.: Harvard University Press, 1971), 127.

[3] Michael Walzer, "The Obligations of Oppressed Minorities," in *Obligations: Essays on Disobedience, War and Citizenship* (Cambridge, Mass.: Harvard University Press, 1970), 55.

are abstract or other-directed in content. Some self-interested desires – specifically, certain other-directed benevolent desires – might well outweigh them in moral and political value.

2. Rawls on Moral Motivation

In Rawls's later *Political Liberalism*[4], he distinguishes three kinds of desires that enter into the reasonable moral psychology of citizens as free and equal persons in a suitably idealized well-ordered society: object-dependent, principle-dependent, and conception-dependent desires. These are, respectively, desires which take states of affairs as their objects, those which take principles as their objects, and those which take as their objects the conceptions or ideals that hierarchically ordered sets of principles constitute and articulate. These amplify the moral powers of such citizens – the capacity for a sense of justice and for a conception of the good, and include the intellectual powers of judgment, thought and inference necessary to exercise the moral powers. (81) All three cut across his earlier distinction between interests in and of a self, in that all three can take either self- or other-directed objects. In what follows I focus on the first two, since the implications for my concerns about the third then follow straightforwardly.

2.1. Object-Dependent Desires

Object-dependent desires are those, Rawls says, which can be described without invoking moral or rational principles or concepts. Examples would include bodily desires for food, sleep, or pleasure; socially conditioned desires for wealth, status, power, glory, or property; as well as emotional attachments, loyalties and devotions, and vocational commitments. (82) Rawls thinks these sorts of desires are some of those which a person can have whether or not he understands the principles that identify them. Rawls might say that an agent can desire such objects unselfconsciously, i.e. without theorizing about them to himself. Most of the desires he lists are straightforwardly self-directed. But among these, the last four mentioned – emotional attachments, loyalties and devotions, and vocational commitments – are most flexible in their orientation toward self or other. For example, my emotional attachment to you is other-directed if it is based in disinterested admiration and affection for your personal qualities. It is self-directed if it is based in your physical likeness to me, or your success in satisfying my need for a father figure, or your practice of flattering my vanity. Similarly, your loyalty and devotion to me are other-directed if they are based in your perception of me as trustworthy and kind. They are self-directed if they are based in your expectation of receiving a tidy inheritance in recompense for

[4] John Rawls, *Political Liberalism* (New York: Columbia University Press, 1996). All page references to this work are paginated in brackets in the text.

your attentions, or your fear of my revenge should you betray me. Finally, your vocational desire to become a lawyer is other-directed if you believe lawyers are guardians of justice, or that studying the law is an intrinsically honorable and worthwhile activity. It is self-directed if you believe lawyers have social cachet, or can pull down six-figure salaries during the first year out of law school.

In Chapter II.2, I offered an analysis of desire that questioned whether any desire could be conscious without being conceptualized in some minimal sense. Rawls does not deny that here. When he identifies such desires as object-dependent and describable independent of any moral or rational conception or principle, he means rather to exclude such desires from any necessary dependence on normative criteria of evaluation, i.e. on principles and conceptions relative to which such desires might be interpreted, criticized or thought worthy of revision. Such desires are object-dependent in the sense that it is the envisioned state of affairs that motivates us to achieve it, rather than our attachment to any principle or conception that state of affairs may or may not instantiate.

2.2. Principle-Dependent Desires

Rawls defines principle-dependent desires as those which are

> (1) dependent on the principle in question for an accurate description of the object or end we desire to realize (82);
>
> (2) such that the "force, or weight" of the desire is a function not of its psychological strength, but rather of the principle needed to describe it (82-3, fn. 31);
>
> (3) such that the evaluative priority of the desire is similarly "given entirely by the principle to which the desire is attached, and not by the psychological strength of the desire itself;" and
>
> (4) are of a kind that only a rational or reasonable agent who can understand and apply such principles can have (82).

In this definition, (2) and (3) above seem to be equivalent, i.e. the motivational force of the desire is a function of its evaluative priority in the agent's ordinal ranking. A desire to do the right thing, or to discharge one's responsibilities efficiently, or to preserve one's integrity, or to conserve energy, or to advance the common good, might exemplify principle-dependent desires.

He distinguishes two kinds of principle-dependent desires, depending on the kind of principle invoked to describe them: *rational* principles are those denoted by what I have described as the egocentric principles of rationality that characterize the Humean conception of the self. He identifies as *reasonable* those which "regulate how a plurality of agents (or a community or society of agents), whether of individual persons or groups, are to conduct themselves

in their relations with one another." (83) Rawls's notion of reasonable principles as providing rules of interpersonal coordination would count as a subset of what I have described as principles of transpersonal rationality that characterize the Kantian conception of the self I elaborate in Volume II.

However, Rawls's assumptions about what a Kantian conception of the self requires is quite at odds with the one I have so far sketched only in outline. He claims that

> (A) A person with a good will, to use Kant's term, is someone whose principle-dependent desires have strengths in complete accordance with the force, or priority, of the principles to which they are attached. (83, fn. 31)

Let us first examine Rawls's characterization, before we consider whether it makes any sense to ascribe such a characterization to Kant. As we see, in analyzing principle-dependent desires, Rawls uses the terms force, weight and priority more or less interchangeably, and contrasts them with psychological strength. The former, he claims, provides the important quantitative measure of the motivational strength and evaluative status of the desire, to which its psychological strength is irrelevant. He does not claim that principle-dependent desires have no psychological strength; on the contrary.

> (B) This strength I assume to exist and it may enter into explanations of how people in fact behave but it can never enter into how they should behave, or should have behaved, morally speaking. (83, fn. 31)

So in passage (A) quoted above, his assertion that to have a good will is to have principle-dependent desires whose "strengths [are] in complete accord with the force, or priority, of the principles to which they are attached" must be interpreted as expressing the thought that the principles to which the desire is attached are the measure of its *motivational* strength; for he has just claimed in passage (B), which directly precedes it, that its *psychological* strength is irrelevant to morally obligatory action. In the idealized conception of the well-ordered society to which the reasonable moral psychology of its citizens is pertinent, only how people should behave, morally speaking, is to be considered; considerations of the psychological strength of various motives play no role. Rawls's thesis, then, is that such citizens are motivated by rational and reasonable principle-dependent desires to act in accordance with those principles whose priority, or weight, or force, is strongest under a given set of circumstances, regardless of their psychological strength. The basic idea is that ideally, my principle-dependent desire to, for example, do the right thing, or conserve energy, takes motivational priority over self-serving or wasteful object-dependent desires that may have greater psychological strength for me.

On the face of it, it is difficult to understand the distinction Rawls seems to want to draw between psychological and motivational strength. If the strength of my motivation to do something is not to be understood

psychologically, and does not enter into my psychological state, how does it manage to propel me into action? But if it does not manage to propel me into action, what is the point of talking about it? On the other hand, if, as one might hope, my motivation does after all affect me psychologically, then in what respect is motivational strength inherently different from psychological strength? Why distinguish between them at all? Why not just say that we often have many different desires, of different psychological strengths, some of which are greater and some less, such that the strongest is motivationally overriding and such that principle-dependent desires can occupy that role? For unless there is some more comprehensive internal principle that weighs and evaluates psychological and motivational strength relative to each other, we are then susceptible to making some rather peculiar decisions as to what to do.

Thus suppose I have a principle-dependent, rational desire to maximize utility and I determine, after careful reflection, that this requires maximizing my personal wealth (thus this is not an object-dependent desire for wealth itself). Suppose that on this basis, I ascribe higher priority to being a rich dentist than to being a poet (thus being a rich dentist is similarly not an object-dependent desire in my scheme of things, but rather embedded in my principle-dependent desire to maximize utility). My desire to become a rich, utility-maximizing dentist satisfies Rawls' fourfold definition of a principle-dependent desire, in that

> (1′) it is dependent on the utility-maximization principle for an accurate description of the end I desire to realize by becoming a rich, utility-maximizing dentist;
>
> (2′) the "force, or weight" of my desire in my moral psychology is a function not of its psychological strength, but rather of the utility-maximization principle needed to properly describe it;
>
> (3′) the evaluative priority of my desire to become a rich, utility-maximizing dentist is similarly given entirely by the utility-maximization principle to which that desire is attached, rather than by its psychological strength; and
>
> (4′) the desire is of a kind that only a rational agent who can understand and apply the utility-maximization principle can have.

Suppose that I then devote myself to the vocational end of becoming a rich, utility-maximizing dentist, even though being a poet has greater psychological strength for me, and even though I have no object-dependent desire to be a rich, utility-maximizing dentist. That my principle-dependent desire to be a rich, utility-maximizing dentist should override my object-dependent desire to be a poet seems counterintuitive at best, psychologically unhealthy at worst.

Next consider my principle-dependent, reasonable desire to tell the truth, such that I determine, again after careful reflection, that this requires criticizing my supervisor's fashion choices, despite the dangers to my continued and future employment. Again a review of Rawls's fourfold definition of a principle-dependent desire will show that this desire satisfies its criteria. Then suppose that on this basis I ascribe higher priority to truthfully criticizing my supervisor's fashion choices, and so boldly speak out on this score, even though securing my continued and future employment has greater psychological strength for me, and even though I have no object-dependent devotion to my supervisor's sartorial self-improvement. Again this commitment to my higher-priority, principle-dependent desire at the expense of a psychologically stronger, object-dependent one that violates it seems misguided.

What is wrong, in both cases, is that an object-dependent desire of greatest psychological strength is subordinated to a principle-dependent desire having strongest normative priority, even though commonsensical rationality would seem to rest with the desire thus subordinated: If I want to be a poet I should be one, even though satisfying this object-dependent desire violates my principle-dependent desire to maximize utility. Similarly, if I want to keep my job I should murmur, politely but unintelligibly, in response to my supervisor's bright question, "Well, how do I look?" – even though this object-dependent desire is incompatible with my higher-priority principle-dependent desire to tell the truth. What is lacking – for Rawls here as well as for the other Humeans to be considered in Chapter VIII – is a higher-order principle of rationality that would enable us to adjudicate sensibly between these conflicting desires.

What is more deeply wrong, however, is Rawls's separation of psychological strength from normative priority in the first place. He wants to claim that a consideration having highest normative priority can thereby have greatest motivational strength independent of its psychological strength. He wants to reserve the psychological strength of a desire for *de facto* causal explanations of motivation, in which agents are by definition moved by that desire that has the greatest psychological strength for them at that moment, as the Humean belief-desire model requires. Normative priority (or weight, or force), on the other hand, is supposed to determine the relative status of the desire in the agent's ideal ordinal ranking, such that the highest normative priority (or weight, or force) of a desire under particular circumstances is a necessary and sufficient condition of that desire's counting as the best all-things-considered reason for acting. Greatest psychological strength can diverge from highest normative priority, because the desires on which agents do in fact act can diverge from the desires that give them most reason to act. Rawls attempts to close this gap by ascribing greatest motivational strength to that desire with highest normative priority, independent of its psychological

strength. Rawls's idealization stipulates that citizens of the well-ordered society always act on those desires that most give them reason to act, regardless of the psychological strength of any desires that may or may not conflict with them.

But this merely rehearses the Humean externalist's strategy against which – as we see in the next chapter – Thomas Nagel fought so hard. Humean externalists need a distinction between psychological strength and normative priority, in order to explain how a desire can be recognizably rational yet fail to inspire one to act on it. Rawls's idealized moral psychology does not need a separate and mysterious concept of motivational strength to close the gap between reason and action, because in the well-ordered society, there is no reason why normative priority should not directly determine psychological strength in the moral psychology of its citizens. What he should have said was simply that in citizens of the well-ordered society, principle-dependent desires with the highest normative priority under given circumstances therefore have the greatest psychological strength as well, and that such citizens therefore act without conflict to satisfy such desires.

2.3. Rawls versus Kant

What Rawls should have said about principle-dependent desires is similar to what Kant does say about principles dependent on reason, when he speaks to the difference between how we are actually motivated and how we ideally would be motivated:

> A perfectly good will would thus stand quite as much under objective laws (laws of the good), but it could not on this account be conceive as necessitated to act in conformity with law … 'I ought' is here out of place, because 'I will' is already of itself necessarily in harmony with the law. (G, Ac. 414) … for this 'I ought' is properly an 'I will' which holds necessarily for every rational being – provided that reason in him is practical without any hindrance. (G, Ac. 449)[5]

These two passages from the *Groundwork* assume what Rawls in passage 2.2.(B) above denies, that the psychological strength of one's rational motive can enter into how one should behave, morally speaking. Kant speaks to the psychological strength of one's rational motive by noting that the factor of necessitation, the sense of obligation or duty to act in accordance with the moral law, is absent in the ideal case. Ideally, when reason motivates us "without any hindrance," our psychological state is one of "harmony with the law," and reason itself has the greatest psychological strength. In this case, it is moral obligation – "oughts" and "shoulds" – that is irrelevant to Kant's

[5] Kant, Immanuel, *Groundwork of the Metaphysic of Morals*, trans. H. J. Paton (New York, NY: Harper Torchbooks, 1964). Academy Edition reference to this work are paginated in the text, preceded by "G".

ideal. Whereas Kant's is an ideal of motivational harmony between psychological strength and rational requirement, Rawls's is an ideal of motivationally effective moral obligation. We can think of moral theorizing about action as including three levels of idealization: first, the non-ideal reality, in which desire often runs rampant; second, the ideal of desire subordinate to the demands of moral obligation; and third, the ideal of desire in spontaneous rational harmony with moral principle. These correspond very roughly to Aristotle's typology of the *akratic*, the *enkratic*, and the *agathos*. In this hierarchy, Rawls's conception of moral motivation occupies the second level of idealization, whereas Kant's perfectly good will occupies the third.

Moreover, these two passages from the *Groundwork* deny what Rawls in passage 2.2.(A) above assumes, that this rational motive for the perfectly good will might be equated with desire rather than reason. For Kant, "reason … is practical without any hindrance" when it is without the hindrance of *desire* (unless of course reason is to be equated with desire – an unprecedented and even more radical extension of the concept of desire into vacuity that Rawls would surely reject). Kant's ideal of moral motivation is one in which principles of reason themselves have all the motivational force and psychological strength needed to effect the required action. For Kant, it is the tension between reason and desire that engenders the sense of duty, the "necessitation" that marks moral obligation. Kant's assertion that moral obligation is irrelevant in the ideal case therefore entails that desire is, too.

Thus Rawls's invocation of Kant's notion of the good will is, in effect, a *non sequitur*. Kant's question is basically the one with which I began this project in Chapter I, as to whether or not we have the capacity to act according to the dictates of reason, even when this conflicts with our personal desires and interests. He answers in the affirmative, by arguing that reason itself, independent of desire, can indeed be motivationally effective. Kant's solution to the problem of moral motivation thus stipulates two sources of motivation within the self, reason and desire; and requires a convincing argument that reason can outcompete desire in moving the agent to action. By contrast, Rawls assumes from the outset not only that Kant's argument does not convince, but furthermore that no revision of it that respects Kant's bipartite conception can.

Now Rawls takes pains to emphasize the "obvious non-Humean character of this account." (84)[6] He justifies his description of his account as "non-Humean" on the grounds that

[6] I take it that Rawls in this section of *Political Liberalism* means to, among other things, defend himself against my description of him as a Humean in my "Instrumentalism, Objectivity, and Moral Justification" (*American Philosophical Quarterly 23,* 4 (October 1986), 373-381). However, it is difficult to tell, as Rawls does not cite this or any other papers of mine in the text, even when his words on the page echo their theses with near-perfect fidelity. I believe I make clear in both places my definition of the term,

> it runs counter to attempts to limit the kinds of motives people may have. Once we grant – what seems plainly true – that there exist principle-dependent and conception-dependent desires, along with desires to realize various political and moral ideals, then the class of possible motives is wide open. … How is one to fix limits on what people might be moved by in thought and deliberation and hence may act from? (84-85)

How, indeed? Precisely the problem we repeatedly encounter with the Humean belief-desire model of motivation is that it fixes no limits whatsoever on what can count as a desire, and therefore no limits that might enable us to distinguish meaningfully between desire and any other type of motive. And so one answer to Rawls's rhetorical question here might be simply to allow people to be moved by thought and deliberation themselves, while limiting the conceptual reach of desire to the nonvacuous. I develop this answer in Volume II.

Rawls thus implicitly accepts the Humean conception of the self as authoritative, despite his protestations; and with it the Humean model of motivation that stipulates desire as the sole explanatory variable. Rawls contents himself with offering subtle distinctions in the types of desire on offer. These distinctions are useful. There certainly is a difference between wanting a thing or state of affairs, wanting to conform one's actions to a principle, and wanting to conform one's actions to principles that define an idealized self-conception or social conception. Nevertheless it is true of all of these different types of desire that they are, at the end of the day, desires; i.e. wants – represented lacks that the agent acts to replenish. Since Kant's ideal of the perfectly good will already does, always and necessarily, act in spontaneous harmony with principles of reason, it is in theory and by definition impossible for the perfectly good will to be motivated by any such want.

So Rawls's account of moral motivation does not escape the problem of moral motivation that besets all Humean accounts, because he is irrevocably committed to desire as the sole motivation of action. His three-fold distinction among object-dependent, principle-dependent, and conception-dependent desires does not alter this commitment. And so it is true of all such desires, as for all the others, that if I happen to lack such a desire in the non-ideal case, I

"Humean;" and make it even clearer in my "Two Conceptions of the Self," (*Philosophical Studies 48*, 2 (September l985), 173-197; reprinted in *The Philosopher's Annual VIII* (1985), 222-246), drawn from the dissertation I wrote under Rawls' supervision. Rawls' treatment of desire in all of his works, starting with *A Theory of Justice,* squares nicely with this definition. If Rawls really means to "dissolve the line between [Williams'] allegedly Humean view of motivation and Kant's view, or ones related to it" (85, fn. 33) as he claims, then he should have no objection to being called either a Humean or a Kantian indifferently.

then have no motivation to achieve the end in question; that I do have motivation to achieve the end in question only if doing so brings me some measure of personal satisfaction and not otherwise; and that therefore, desiring to achieve the end in question, whether self-directed or other-directed and whether object-, principle-, or conception-dependent, is a self-interested motive. Moral motivation, as we ordinarily understand that term, seems just as out of reach on Rawls's account of it as it does for that of any other committed Humean.

3. Desire-Satisfaction and Personal Gain

3.1. Pleasure

However, from the thesis that all desires, including other-directed and principle-dependent ones, entail anticipated personal satisfaction, it does not follow that all desire-satisfaction entails pleasure. We have just seen in Section 1, and more fully in Chapter II.2 above, that the satisfaction of a desire is the provision of something experienced as lacking – literally, wanting. The experience of satisfaction is the experience of sufficiency restored. One may satisfy a desire without obtaining pleasure – even a small one – if the satisfaction of that desire instead causes one boredom, or discontent, or pain. In this case what was wanting has been supplied, thereby restoring sufficiency – only to cause, in turn, further things to be wanting: interest, or contentment, or the cessation of pain.

An example of a desire whose satisfaction might cause one boredom rather than pleasure is the desire not to waiver in one's morning routine: ablutions, calisthenics, walk the dog, boiled egg and coffee, off to work. In this case the restoration of sufficiency involved in adhering to one's morning routine creates monotony: a want of – i.e. for – stimulation and variety. An example of a desire whose satisfaction might cause one discontent rather than pleasure is the final seduction of a distant and longstanding crush, whose proximity and detail destroy the romantic illusion of perfection. How can the desire-satisfaction *itself* cause dissatisfaction? A short answer is that the object as desired – i.e. the intentional object – is not the same as the actual object to which the desire refers, and it is in measuring the gap between them that insufficiency is to be found. A longer answer was offered in Chapter II.

An example of a desire whose satisfaction might cause one pain rather than pleasure is the desire for public recognition, which may bring envy, enmity, betrayals, and harassment in its wake. Here the achievement of the object of desire creates the further desires for friendship, trustworthiness, and privacy – all of which are now wanting precisely and only because the original object of desire no longer is. In all of these cases, we take an interest in satisfying our desires, whether self- or other-directed, because of the personal

gain in satisfaction we thereby obtain, even where we may neither value nor focus on that satisfaction.

Satisfying our own desires, then, can be an object of interest for us – i.e. it can be an interest we take in the condition of our selves, even when it is not itself an object of pleasure, or even itself an object of desire – i.e. even when we do not envision the prospect of satisfying our desires as itself satisfying. For example, we might be disgusted or embarrassed by our desires, and envision their satisfaction with distaste or horror: we visualize ourselves at the moment of satisfaction, collapsing, weak-limbed, under the intense pleasure of long-deferred gratification; and realize that we are, in this state, not only abject and debased but also ridiculous.

3.2. Self-Direction vs. Self-Interest

Classifying other-directed desire-satisfaction as an instance of personal gain and self-interest because of the anticipated personal satisfaction it entails does not in turn imply a view popular in some contemporary psychotherapy circles. According to this view, a "co-dependent" is a person, motivated by other-directed desires for another's well-being, to sacrifice his health, peace of mind, and/or financial security to care for that other, in order to obtain a sense of control, power, or self-esteem. This analysis treats ostensibly other-directed behavior as not only self-interested but also ultimately self-*directed*, i.e. as fueled by a quest for personal control, power, or self-worth. It identifies the desire for another's well-being as instrumental to the satisfaction of an ultimate and motivating desire for personal control, etc. It thus reduces apparent altruism to actual egoism. I do not doubt that this analysis holds true for certain personalities under certain circumstances. But it would not be plausible to generalize this analysis into full-blown Psychological Egoism, i.e. into a thesis that all agents at all times, regardless of the ostensible content of their ends, are ultimately motivated by such desires for personal advantage, as Hobbes tried to do (on pain of contracting all the elementary theoretical objections to which Psychological Egoism is subject).

By contrast, the thesis that all desire is a species of self-interested motivation can be so generalized, because it is entailed by a sharp distinction between desire and interest, plus a conceptual analysis of what a desire is. To summarize very briefly the analysis offered in Chapter II.2.1, a *desire*, regardless of content, is a motivationally effective psychological state whose object is envisioned as a source of personal satisfaction, such that the envisioned satisfaction of the desire is what moves one to achieve it. A Humean might be tempted to object that in addition to the satisfaction, surely the content of the object of desire itself, independently, also plays a role in moving us to achieve it. But it is only the fact that this content is part of the object of *desire*, and so conceived as wanting by the agent, that confers any special conative power on it. Aside from the agent's envisioning of the object

as satisfying a lack, nothing about the object considered independently of its status as an object of desire might motivate the agent to realize it. At least not within the constraints of the belief-desire model of motivation strictly understood. A significant modification of this model is considered in Chapter XI.

I have already pointed out that one can envision an object as satisfying without having the desire to achieve it. To this can be added that in fact, one can envision an object as satisfying although it is the intensional object of a completely different kind of motivation, such as an intention or resolution. For example, I might resolve to organize my time more efficiently, envision the resulting state of order and control as deeply satisfying, yet be motivated, not by the desire for that satisfying experience of order and control, but rather by that very resolve, which in turn is driven by the conviction that efficient time-management is a necessary condition of personal autonomy – which I experience, not as a source of satisfaction, but rather as a source of self-realization (self-realization is not necessarily satisfying because the aspects of the self one realizes may be deeply troubling – or trouble-making). Of course the belief-desire model of motivation would deny the accuracy of this description, and claim instead that it denoted a *desire* for self-realization in disguise.

The thesis that all desire is a species of self-interested motivation does not distinguish between the actual and ostensible content of a desire, as the Freudian variant on the belief-desire model would, because this distinction does not require any modification in the basic definition of a desire. The Freudian variant is thus merely a special case of the more general thesis. Nor does this thesis then go on to claim, as the Freudian variant would, that all actual desire is ultimately self-directed, for the Freudian variant is in fact merely a subvariant of the Hobbesian one, with all of its attendant problems. Instead, this thesis respects the distinction between self- and other-directed desires all the way down. The thesis is simply that if it is indeed a *desire* that motivates one, then regardless of the content of that desire, one anticipates personal satisfaction from obtaining its object. This is just a roundabout way of elaborating on the truism that we are motivated to *satisfy* our desires, and that we are motivated to satisfy *our* desires.

3.3. Criterion-Satisfaction

Humeans who dislike the moral implications of this thesis may complain that I have phenomenalized to excess the concept of desire-satisfaction. Desire-satisfaction, they may argue, is less like the chops-licking, warm-glow phenomenal sense of satisfaction-as-fullness one experiences after a good meal, and more like the satisfaction of a criterion. That is, like criterion-satisfaction, desire-satisfaction merely supplies in reality a condition or state of affairs required by the desire itself. Understood in this sense, satisfaction

does not imply personal gain. Therefore, Humeans may argue, there are no grounds for grouping other-directed desire-satisfaction under the rubric of self-interest.

But first, it is not obvious that this much weaker interpretation of desire-satisfaction conclusively rules out the implication of personal gain, because there are cases in which one may gain personally from the satisfaction of a criterion. For example, a job applicant may gain – or lose – from satisfaction of the unstated criterion governing a job search that the successful applicant look and behave just like everyone else in the organization. In such cases, desire-satisfaction (here, satisfaction of one's desire to be hired) is a special case of criterion-satisfaction more generally (here, satisfaction of the criteria a successful job applicant must meet).

A fortiori, just because the satisfaction of my self-directed desire for a rich and anonymous benefactor contains no phenomenal component (let us suppose I believe, rather, that I am simply receiving a surprisingly high rate of return on my investments), this does not entail that I receive no personal gain. My anonymous benefactor's charitable contributions to my money market account are my personal gain, regardless of what I believe about where they come from. Nor does the absence of a phenomenal component in the satisfaction of my other-directed desire for a rich and anonymous benefactor for my best friend (here I assume rather that *she* is merely doing unusually well with her investments) entail that I receive no personal gain. My personal gain here is her increased income, even if I am mistaken about its origins and ignorant that this other-directed desire of mine is, in fact, being satisfied. I can receive personal gain without being phenomenally aware of it, as when I inherit a rental property without knowing it. Whether I am aware of it or not, if I get what I want – literally, what I conceive myself to lack – then I gain what I have gotten, even if it is the thing gotten rather than the fact of my getting it that holds my attention. Therefore even other-directed desire-satisfaction is a species of self-interested motivation.

But in any case, thirdly, it won't do simply to deny that desire-satisfaction is phenomenal and replace it with an imaginative conceptual analogy, as this argument tries to do. This begs the question. If the satisfaction of a compulsion, appetite, or craving for a good meal include a phenomenal component, it is difficult to understand why a desire for one would not. And if the satisfaction of a desire for a good meal includes a phenomenal component, it is even harder to see why the satisfaction of desires for many other things should not.

Finally, the imaginative conceptual analogy itself as stated fails to distinguish among different psychological "criteria" which the supplied object is supposed to "satisfy". The object or state of affairs that satisfies the *desire* for *x* may, but need not be, identical with that which satisfies the *resolve* to do *y*, the *will* to *z*, the *intention* to do *w*, or the *craving* for *r*. On this weaker

interpretation of "satisfaction", the satisfaction-relation that holds between the desire, resolve, will, intention, and craving, and the objects *x*, *y*, *z*, *w* and *r* respectively that are supposed to satisfy each intentional state as enumerated would be the same. But the satisfaction-relation is *not* exactly the same for each: *y* and *w* must be an action, whereas *x* and *z* need not be and *r* is highly unlikely to be. If the stipulated satisfaction-relation cannot distinguish between the requirements of a desire, a resolve, a will, an intention, and a craving, what good is it?

4. Malevolent Other-Directed Desires

Like self-directed desires, other-directed desires can be selfish or self-indulgent without being ultimately self-directed. An other-directed desire is *selfish* if the agent accords more value to the gain its satisfaction brings him than to the gain its satisfaction brings to the others to whom it is directed. This is consistent with the object of the desire's being the other in question rather than the concomitant experience of satisfaction that outweighs it in value. Any malevolent or spiteful desire would serve as an example, since in these cases the value of the gain its satisfaction brings oneself is inversely proportional to the value of the gain its satisfaction brings to the other (and directly proportional to the value of the other's loss): If you want me to fail in my ambitions, then the less satisfaction I obtain from my strivings, the more you obtain by thwarting them, and the more I fail, the better you like it.[7]

But benevolent other-directed desires can be selfish, too, if the agent accords higher value to the gain its satisfaction brings her than to the gain its satisfaction brings the other to whom it is directed. Someone who takes great satisfaction in charitable fundraising but is comparatively indifferent to the gains this activity will entail for its recipients would be an example. Here selfishness might manifest itself as deep-seated frustration or resistance to the discovery that its recipients are harmed rather than helped by it.

An other-directed desire is *self-indulgent* if its satisfaction is impulsive and self-destructive to the agent, even if it is beneficial to the other at whom it is directed. For example, someone who satisfies five times a day his impulsive desire to call his partner at work may be indulging a caring impulse that may gradually undermine his autonomy, even if it is directed at making his partner feel loved and succeeds at doing so.

4.1. Brutalization

Like benevolent desires, malevolent other-directed desires can be self-indulgent if their satisfaction is not only self-destructive but also impulsive. I shall refer to malevolent other-directed desires, i.e. desires to deliberately

[7]On the existence of bona fide malevolent desires, see Michael Stocker, "Desiring the Bad: An Essay in Moral Psychology," *The Journal of Philosophy LXXVI*, 12 (December 1979), 738-753.

inflict harm on others, as *sadistic*. A sadistic person finds satisfaction, not merely in others' suffering – such a person would be more properly described as spiteful or *schadenfroh*; but rather in actively inflicting that suffering on others. A sadistic person takes satisfaction both in the other's suffering itself; and, just as important, in being the instrument of that suffering. Michael Slote argues that sadism, like other inherently vicious pleasures, such as drug addiction, as well as wealth and power, may be personal goods for the virtuous agent who has them, even if they violate the constraints of morality and therefore provide no reason for the virtuous agent to act.[8]

Against Slote's view, I contend that satisfying sadistic desires is always self-destructive, and at least as destructive as the experience of externally inflicted harm on its victims. Consider first the latter case. Through externally inflicted harm, its victims thereby accumulate experiences and memories of aggression directed against the self, and these experiences and memories in turn have a harmful conditioning effect on the integrity of the self. They disrupt the equilibrium and coherence of the self by disrupting the equilibrium and coherence of the external order the self experiences, and replacing it with affective images of aggressively inflicted pain, violence, or disorder. Just as we speak of the corruption of a text that is rendered unsound and tainted by external interpolations and emendations, we may speak similarly of the corruption of a self that is rendered unsound and tainted by the interpolation of external, destabilizing and disruptive experiences of aggression directed against it.

The more numerous and familiar these corruptive experiences become, the more they vitiate the equation of well-being with stability and order, and the more they desensitize the self to the danger they represent to its stability and integrity. Aggression repeatedly directed against its victim habituates the victim's self to a condition of disintegrity, and so to a lack of interest in the self- or other-destructive consequences of its behavior. There is thus a continuum of damaging consequences to the self of externally inflicted harm – and a corresponding continuum of corrigibility – with simple insensitivity at one end, and pervasive and uncontrolled brutality at the other.

Individuals who are *brutalized* by the violence or abuse they experience at others' hands are – by definition – more capable of inflicting similar violence or abuse on others in turn. Indeed, whenever one witnesses another person's brutality, whether physical or psychological, it can be useful to ask oneself where and from whom the other learned to behave that way. It would be a mistake to think of brutalization as a process requiring physical violence. The self can be brutalized in more subtle ways through verbal, emotional and psychological abuse and manipulation as well, which in turn may cause

[8] Michael Slote, *Goods and Virtues* (New York: Oxford University Press, 1983), Chapter V.

physical harm. Associating or fraternizing with brutal people can have a similarly corruptive effect.

4.2. Sadism and Self-Brutalization

Actively inflicting harm on others – and so satisfying sadistic desires – is thus an expression of brutality that accelerates this brutalizing process even more. Just as with externally inflicted harm on oneself, inflicting harm on others accumulates in one experiences and memories of aggression and harm, and these experiences and memories have a similarly harmful conditioning effect on the integrity of the self. They similarly disrupt the equilibrium and coherence of the self by disrupting the equilibrium and coherence of the external order the self experiences, and replacing it with affective images of pain, violence, or disorder. And just as with externally inflicted harm, the more numerous and familiar these corruptive experiences become, the more they vitiate the equation of well-being with stability and order, and the more they desensitize one to the danger they represent to the stability and integrity of one's self.

But to these already brutalizing effects, inflicting harm on others adds *self-brutalization*: the infliction on oneself of the experience of inflicting harm on others. Quite aside from the effect of any moral emotions such as guilt, shame, remorse, or self-dislike, or justified retributions one may or may not experience as the result of having inflicted harm on others, the experience of inflicting that harm further desensitizes one to the self- and other-destructive consequences of one's behavior. Unlike the experience of being harmed, the experience of inflicting harm on others is actively self-initiated. It habituates one, not only to the experience of external aggression directed against one's self, but to the experience of actively directing that aggression against the external world that provides the conditions of coherence of one's self. To inflict harm on the world as one views it is thereby to inflict on one's self the same harmful experiences and memories of pain, violence, and disorder. One further desensitizes oneself as one further habituates oneself to these experiences and memories, by originating, performing, and repeating them. Since one's own infliction of harm on others is itself an experience that issues from and reinforces the sense of self one already has, further self-brutalization is inevitable. So satisfying sadistic desires is finally more destructive of the self than being their victim, because it absorbs, reinforces, and accelerates the brutalizing process of corruption, destabilization, and desensitization of the self to external factors that destroy its integrity.

4.3. Malice

But in addition to being necessarily self-destructive, satisfying malevolent other-directed desires can also be impulsive rather than deliberate. Such desires are not sadistic but rather *malicious*. Lying, hypocrisy, spiteful gossip,

verbal, emotional, or physical abuse, battering, or destruction of persons or property are just a few on the continuum of impulsive malevolent acts that may both be unpremeditated and also satisfy a malevolent other-directed desire. Where these acts are genuinely *im*pulsive rather than *com*pulsive, they are also self-indulgent – though their malevolence may overshadow their self-indulgence in our judgment of them. Malicious desires always include a self-indulgent component, because their satisfaction is impulsive and self-destructive to the agent for the reasons mentioned above.

Like benevolent desires, malevolent desires can also be genuinely other-directed without being ultimately self-directed. So, for example, one may have a straightforwardly sadistic or malicious desire to inflict harm on another, and receive personal satisfaction from doing so, without the object of that desire's being conceived as a means, either consciously or unconsciously, to a sense of power, control, or self-esteem in oneself. In fact, one may gain satisfaction from inflicting harm on the other at the same time that this satisfaction undermines one's sense of power, control, or self-esteem. So, for example, the feeling of satisfaction may diminish one's self-esteem if it illuminates too clearly for comfort one's distasteful motives. Or the action that causes this feeling – infliction of the actual harm the other experiences – may reduce one's sense of control by surprising one with the full extent and force of one's cruelty. In any such case, the object of one's desire may well be inflicting the harm rather than the personal gain in satisfaction one obtains by inflicting it. That one anticipates obtaining this personal gain is what makes satisfying a malevolent other-directed desire a species of self-interested motivation nevertheless.

5. Desire-Satisfaction and the Moral Interests of a Self

5.1. Moral Considerations

Now if one is sure that we are motivated to satisfy our desires, whatever they are, and doubts that we are motivated by anything else, then *a fortiori* one doubts that moral considerations alone can motivate us to act in others' interests. From this perspective there are two things wrong with moral considerations alone: First, they are not desires; and second, therefore, they cannot be brought under the rubric of self-interested motivation, whether immediate or long-term, although they may coincide with either. We have seen in Section 1.3 that moral considerations alone would be among what Rawls calls the interests of a self; and may call upon us not only to disregard, but maybe even to sacrifice completely our interests in our selves – including our interest in satisfying both self- and other-directed desires.

For example, moral considerations alone may require a parent to ignore her benevolent other-directed desire to provide for the material security in perpetuity of her children, on the grounds that this will stunt their capacity

for growth, independence, and initiative. A parent may be secretly attracted to the idea of thus reining in the ungrateful little squirts; of securing her children's best interests despite their precarious and (to her) ill-considered lifestyle predilections – drugs, tattoos, scarification, pierced lips, and the like. Yet purely moral considerations may require her to let them sink or swim on their own, and suffer the consequent anxiety and concern in silence. Alternately, moral considerations alone may require us to sacrifice our self-directed desire for material security in perpetuity for ourselves, on the grounds that it is excessive given the scarcity of material and social resources available over all. Indeed, moral considerations alone may permit us to heed our self-or other-directed desires only under the most limited and innocuous circumstances.

To think it is psychologically impossible for us to observe any of these strictures without a desire to do so, is to accept the familiar Humean, belief-desire model of motivation, according to which all action is motivated by the agent's desire to realize a certain object or state of affairs and the agent performs that action he believes will most efficiently achieve this goal. This model of motivation implies that rational appeals, argument and dialogue are *in theory* insufficient to reform attitudes, change minds, create desires, or inspire action because, on the Humean model, desire is the only motivationally effective cause of action there can be. Desires can chain-react to prior desires, as instrumental desires are caused by the ultimate desires they effect. And desires may arise from physiological causes, as a drop in blood-sugar level may cause a desire for sweets. But no psychological state other than desire has motivational efficacy on this view.

Hence according to this model, moral dialogue and justification more specifically are equally impotent to reform the culpable. Publications by normative moral philosophers that propose substantive casuistical solutions to such pressing problems as abortion, euthanasia, cloning, or racism are best understood as engaging in abstract philosophical exercises that can have no independent practical import. Similarly, the conferences and symposia on pressing political issues we organize with a sense of the urgency of finding viable solutions reduce to little more than exercises in group self-stimulation and professional networking. This implication will be of little concern to those who regard moral philosophy as nothing but an amusing game. But those who pursue it with an eye to practical re-evaluation and reform have cause for worry if the Humean conception of the self is the correct one.

5.2. Whistle-Blowers, Etc.

There is a problem of moral motivation only on the assumption that the Humean conception of the self *is* the correct one. Obviously, a motivational model that stipulates desire as the only motivationally effective cause of action rules out the possibility of action not motivated by a desire for some

state of affairs; and therefore, I have argued, rules out the possibility of non-self-interested action. But the problem of moral motivation arises because on the face of it there seem to be many such actions – actions we are motivated to perform even though we anticipate not only no satisfaction from their outcome, but even, in some cases, considerable frustration and suffering.

Consider, for example, the whistleblowers of the 1970s and 1980s, before government incentives and protections for federally employed whistleblowers were written into law; as well as those legally unprotected whistleblowers in the private sector up to the present time. This brand of whistleblower is motivated to expose organizational or institutional injustice, and so end it, even though she could have gained everything of value to her by permitting it, and even though she can realistically anticipate losing everything – job, reputation, close relationships, financial well-being, perhaps even her country or her life – by exposing it to public condemnation.

In some such cases, a whistleblower may well take actual satisfaction in seeing justice done or helping others.[9] In these cases, it would be appropriate to describe him as motivated by a desire to see justice done, despite the hardships he suffers in retaliation for his action. But in most cases, the whistleblower may not take any satisfaction in seeing justice done. In some of these other cases, although not all of them, she may at least take *consolation* from seeing justice done. But this is by hypothesis a mere compensation for her losses. It is not the provision of a want – or lack – she has gained.[10] By hypothesis, there are no such personal gains.

Such whistleblowers cite very different explanations for their actions: disgust or outrage with others' arrogance and dishonesty;[11] a belief in open information, truth, justice, or reason;[12] loyalty to the public;[13] conscience or personal ethical or religious principle;[14] a sense of personal responsibility or

---

[9]See Myron Peretz Glazer and Penina Migdal Glazer, *The Whistleblowers: Exposing Corruption in Government and Industry* (New York: Basic Books, 1989), 209-215, 217. Also see Clyde H. Farnsworth, "Survey of Whistle Blowers Finds Retaliation but Few Regrets," *The New York Times* (Sunday, February 22, 1987), page ?.

[10]Not all lacks are losses because some lacks – desires among them – do not necessarily presuppose prior privation. The creation of desire is discussed at greater length in Chapter II.

[11]Glazer and Glazer, *op. cit.* Note 9; page 19, 100, 122, 138, 223, 246. Also see Mary Schiavo, "Flying into Trouble," *Time* (March 31, 1997), pages 52-62.

[12]*ibid*. pages 33, 43, 70, 96, 107. Also see Philip J. Hilts, "Why Whistle-Blowers Can Seem a Little Crazy," *The New York Times* (Sunday, June 13, 1993), Section 4, page 6).

[13]*Ibid*., pages 17, 40, 45, 129.

[14]*Ibid*. pages 43, 70, 88, 96, 101, 103, 104-5, 117, 119, 122, 141, 248-9. Also see Clyde H. Farnsworth, *op. cit*. Note 9; and "In Defense of the Government's Whistle Blowers," *The New York Times* (Tuesday, July 26, 1988), page B6.

obligation to others.[15] Yet they are motivated to blow the whistle anyway – even though they have everything to lose and nothing, not even the satisfaction of a desire, to gain. Indeed, whistleblowers very often express the belief that they had no choice, that they were forced or compelled to expose the corruption of their organizations.[16] Socrates, the most famous whistleblower of them all, offers these considerations in defense of exposing to public ridicule the ignorance and pretentiousness of his fellow citizens:

> Perhaps someone will say: 'Are you not ashamed, Socrates, of leading a life which is very likely now to cause your death?' I should answer him with justice, and say: 'My friend, if you think that a man of any worth at all ought to reckon the chances of life and death when he acts, or that he ought to think of anything but whether he is acting justly or unjustly, and as a good or a bad man would act, you are mistaken.' … Wherever a man's station is, whether he has chosen it of his own will, or whether he has been placed at it by his commander, there it is his duty to remain and face the danger without thinking of death or of any other thing except disgrace. … [I]t would be very strange conduct on my part if I were to desert my station now from fear of death or of any other thing when the god has commanded me – as I am persuaded that he has done – to spend my life in searching for wisdom, and in examining myself and others.[17]

I do not offer a full account of what might motivate a whistleblower to act in these cases until Volume II, Chapter VI.8. But if it is psychologically plausible that a human agent might be motivated by such transpersonally rational considerations to blow the whistle even though she has nothing to gain and everything to lose by doing so – if, that is, there is more to human motivation than can be calculated in a cost-benefit analysis – then there is more to it than can be explained by the belief-desire model of motivation.

It appears that there must be. There is a great deal of ordinary behavior – not only transpersonally rational moral behavior – that we are motivated to perform, not by a desire for their ends, but rather by deeply instilled characterological dispositions to such behavior. Thus, for example, I may regularly dress before leaving the house. I may do so intentionally,

---

[15] *Ibid*., pages 70, 88, 117, 122, 123, 124-5, 129, 130-1. Also see Liz Hunt, "Whistleblowers 'put their health under threat'," *The Independent* (Friday, 10 September 1993), Section 1, p. 6.

[16] *Ibid*., pages 77, 86, 101, 105, 109, 110, 118, 121, 122. Also see N. R. Kleinfeld, "The Whistle Blowers' Morning After," *The New York Times* (Sunday, November 9, 1986), Section 3, page 1; and Don Rosendale, "About Men: A Whistle-Blower," *The New York Times Magazine* (Sunday, June 7, 1987), page 56.

[17] Plato, *Apology* XV.28 – XVII.29, in *Euthyphro, Apology, Crito*, Trans. F. J. Church and Robert D. Cumming (New York: Bobbs-Merrill, 1956), 34-35. I am particularly fond of this translation because its introduction alludes indirectly to the mid-century American political repression of whistleblowers – i.e. McCarthyism – taking place at that time.

deliberately, and consciously. But not because I have any dispositional or occurrent desire to do so. (Suppose, for example, that I live in a singularly tolerant community in a warm climate, and have sufficient power in it so that others will accept uncomplainingly my choice to appear in public unclothed, as did Lyndon Johnson at his Texas ranch.) I may dress before leaving the house out of a socially instilled disposition simply to act on the social principle that people are to appear dressed in public. Here my action is caused by a perception of certain external circumstances that actualizes the disposition in question. I may experience consciously no affective motivational state whatsoever.

Another, transpersonal example: I may contribute time and money to Amnesty International, in order to help restore the civil rights of certain political prisoners. But not because I want to, nor even from any benevolent desire to increase the well-being of the prisoners involved. Indeed, I may know perfectly well that in fact their convictions and attitudes toward life represent values I deplore and would actively discourage if they were in the position to promulgate them. Nevertheless, I may find their torture or imprisonment morally unacceptable, and act to prevent it out of sheer moral indignation that their civil rights are being so severely abridged. Here these feelings constitute an affective motivational state. But they need not make me want to aid them. They may simply make me do so.

In neither case is an intervening desire required to explain my behavior. All that is required is a disposition to act on certain principles – of publicly acceptable self-preservation, or of the inalienability of individual civil rights – that is internalized deeply enough to motivate certain responses under certain circumstances. There are other, more mundane examples: Must I desire to brush my teeth each morning in order to do so? Or to say "Hello?" each time I answer the telephone in order to say it? Evidently not. I just do these things reflexively.

These actions, and others like them, are also relatively unproblematic from the point of view of motivation. Typically, the process by which we are socialized includes instilling a broad range of characterological dispositions to emotion and action deeply enough so that the mere recognition of a situation as requiring a certain kind of behavior suffices to elicit that behavior more or less automatically.[18] Deeply internalized dispositions to such behavior shape

[18]We can think of the social processes by which characterological dispositions are instilled as not unlike the process Aristotle describes in the *Nicomachean Ethics* as *habituation* (Book II; trans. Terence Irwin (Indianapolis: Hackett, 1985). We learn to mimic repeatedly, under similar circumstances, the like behavior of elders or peers with whom we identify, or whose approval we seek. The more frequently we rehearse the behavior and are socially reinforced for doing so, the more natural and reflexive it becomes. Thus I mean to use the word "disposition" here in the narrower, psychological sense that denotes a settled and regular tendency to behave or respond in a certain way

our character, not by making us want to, e.g. say "Hello?" when we answer the phone, but by making us do it.

To say that we may be moved to perform such actions by perceptions, emotions or dispositions that are independent of the action's intentional object is not, of course, to deny that the action must have an intentional object; nor even that we may in some sense "visualize" that object to ourselves prior to performing the action directed at it. It is merely to deny that the impetus to perform such an action must invariably come from a desire or "pro-attitude" (in the revisionist sense examined in Chapter II.1) towards that object. Our dispositions to act and react in certain ways under certain circumstances often impel us to perform intentional actions towards the intentional objects of which we have no evaluative attitudes whatsoever.

Hence it would be a mistake, on this account, to suppose that we could distinguish a bona fide case of action from unintentional behavior according to how deliberate or reflexive it is respectively. The prominent examples of reflexive action we are most likely to identify as such are those that stand out by their social impropriety: thus one "loses one's head," or "goes off half-cocked" in performing some "ill-considered" action. But we should not ignore the large body of reflexive but nevertheless genuinely intentional actions we regularly perform without needing to "consider" them -- because the disposition to perform them has been socially instilled and is therefore socially unremarkable.

Now there are many ways of coming to have such dispositions. The easiest way, when a society's social and legal institutions are in good working order, is to be brought up that way. Then if one accepts those social practices one will require no special desire to behave in accordance with them. Merely the recognition of the situation as being of a certain kind will suffice to elicit the normatively appropriate emotions and behavior. We can describe principles that govern the behavior of each, or most, of the members of a community in this way as *socially operative*.

Consider, for example, the woman who is raised from childhood to be a wife and mother, whatever else she is, in a society that accords high social status to women in this role; and who therefore anticipates, without reluctance, that she will eventually become a wife and a mother. To suppose automatically that she therefore desires to be a wife and a mother does not do justice to the complexity of her feelings. For even supposing she can discern anything like desires beneath such a heavy layer of social conditioning, it is far from obvious that whatever desires she may have necessarily underwrite that conditioning. Nevertheless, she may believe, with good reason, that most women, herself included, should be wives and mothers. She may also be

---

under certain recurrent kinds of circumstances, rather than in the more inclusive sense, that denotes an entity's structural propensity to react in a certain way to certain kinds of causal-counterfactual conditions, even if those conditions never obtain.

moved to satisfy her own parents' desires for grandchildren, and to affirm to them her gratitude for their parenting of her, by becoming a wife and mother herself. She may also believe that she brings special talents to these roles and would fulfill them with outstanding success. In conjunction with her deeply internalized dispositions to view herself as a future wife and mother, these reflective beliefs may motivate her to become one, even though her desires speak against it. Having taken on these roles, she may find them pleasurable and satisfying. But these feelings need not represent the satisfaction of her desire to become a wife and a mother. For there may well have been no such desire.

5.3. Psychological Egoism Again

Unsophisticated Humeans are sometimes inclined to try to explain away counterexamples such as that of the whistleblower, the Amnesty International contributor, or the altruistic wife and mother on grounds of self-deception or unconscious repression: One may *think* one is acting disinterestedly, they contend; but in fact one is always satisfying some desire or other: for approval, perhaps, or martyrdom, or for that truly sublime sensation of self-righteousness. This response trivializes and debases all cases of altruism or principled self-sacrifice, such as that shown by Martin Luther King or Nelson Mandela, by depicting the best of us as no better than our worst motives. It also misrepresents as symptoms of false consciousness actions that could be understood equally well as unique and authentic expressions of our widely underdeveloped capacity for genuinely transpersonal rationality.

Third, this response thereby begs the question. It commits the elementary mistake of turning what was supposed to be a matter of contingent psychological fact into a vacuous conceptual truth (we have already seen in Chapters II and III that there are deep reasons for this). An agent who is motivated by political conviction to devote his life to the eventual liberation of his country, which he fully understands will not occur in his lifetime, can always be *interpreted* as acting, rather, to satisfy some temporally proximate desire or other. But then either there is a fact of the matter about whether desire or conviction is doing the actual motivational work, or else the concept of desire is being applied so broadly that the distinction between desire and other kinds of motivation is lost.

More sophisticated Humeans may wonder why, if not all desire-satisfaction causes pleasure, so that desires may involve satisfaction without pleasure, altruistic actions should not be explained in terms of them. They may then wonder what, if other-directed benevolent desires can exist, remains to be explained. But as we have just seen in most whistleblower cases, not all actions need involve even satisfaction, even of the most minimal kind, to the agent who performs them. An act may be fully intentional, deliberate, and other-directed without providing the agent any satisfaction or pleasure

whatsoever. The belief-desire model of motivation "explains" such actions by either ignoring them or defining them out of existence.

Thus the problem of moral motivation cannot be solved within the constraints of the Humean conception of the self. It cannot be solved by characterizing some desires as impersonal; nor by observing that some desires are other-directed; nor by agreeing that not all desire-satisfaction brings pleasure. These emendations to the belief-desire model of motivation do not solve the problem because they are all consistent with the self-interested nature of all desire. The alternatives are then either to insist on the universally self-interested character of all human motivation – thus defining the possibility of moral motivation out of existence; or else to revise our model of human motivation – and restrict the scope of the belief-desire model accordingly – so as to accommodate its actuality. In the next chapter I examine Thomas Nagel's attempt to navigate between the two.

Chapter VII. Nagel's Internalism

With Chapter VI's account of the Humean problem of moral motivation in hand, I now look more closely at the most sustained attempt so far, within the constraints of the belief-desire model, to solve it; and thereby to meet the challenge posed by cases of seeming altruistic action such as whistleblowers. *Internalism* is the view that moral principle can be a motivationally effective reason for action. It thus would seem to take for granted what Humeans categorically deny: that something other than desire can move us to action. However, most internalists are Humeans. They believe that one's desire to perform an action or advance certain ends by means of it is what provides one with both a motive and a *prima facie* reason for that action. So if one has a benevolent desire, or a desire to be moral, or to act on principle, that desire can be a motivationally effective reason for moral action, provided that such action is deemed likely to satisfy one's desire. Humeans believe, then, that moral principle can motivate us to act – provided that it is the object of a desire so to act.

Externalism, by contrast, is the view that a moral principle may provide one with a reason for action without thereby motivating one to act on it. Externalism may find expression in the familiar response, "Your arguments are convincing, but they don't make me care." Hence where motivation is concerned, the externalist may need to invoke the expectation of reward or punishment – and thus finally appeal to considerations of self-interest, desire, or aversion – in order to impel the agent to do what the balance of moral reasons prescribes. Most externalists – as we have just seen, Rawls among them – purport to be non-Humeans, because Hume stipulated desire and moral sentiment, not principle, as both the motivational and the justificatory basis for moral action. But in fact most externalists are Humeans, because they tend to agree with Hume that only desire can be motivationally effective. Since, they assume, only desires can move one to action, moral principles can provide one with reasons for action even if they provide one with no desire, and therefore no motivation, to act on them. Externalism is a position of desperation, a fallback strategy for preserving the importance of moral principle despite being persuaded by the Humean valorization of desire. It does this by ascribing to moral principle the status of a reason even if it cannot have the status of a motive.

Humean internalists and externalists therefore do not disagree on the essentials. Both believe that only desire is motivationally effective. They differ only on whether something other than a causally effective motive can be a not-necessarily-motivating reason for action. Humean internalists think it cannot, whereas Humean externalists think it can.

Thomas Nagel's *The Possibility of Altruism*[1] is the first and only post-war attempt in the Anglo-American tradition to furnish an extended, genuinely non-Humean internalist account of moral motivation. Several decades after its publication, it remains the unsurpassed modern classic of Kantian rationalist moral psychology. However, it does not carve out a clear and identifiable alternative to the belief-desire model of motivation, nor does it mean to. Nagel's *stated* project is to articulate the rational ethical criteria by which the rationality of desire can be evaluated. What he *actually* does is to attack the premise that desire of any kind must motivate action – without, however, explicating a clear alternative to replace it. But Kantians such as Nagel who in effect accept the Humean model are hard pressed to explain how we can be morally motivated at all in its absence. Without such an alternative, it is unclear how the rational ethical criteria for evaluating desire Nagel develops might, as a precipitating cause of action, occurrently influence the desires we happen to have. In this case externalism looms, and the Humean is free to reassert the primacy of desire as the only plausible candidate for human motivation.

Because Nagel rejects the hypothesis that only occurrent desire-states are motivationally effective in causing action, but declines to supply an alternative account of how actions may be caused, he has only one choice: to modify the belief-desire model so as to accommodate the motivational efficacy of reason within it. Nagel distinguishes between "unmotivated" desires that just assail us, such as appetites, and "motivated" desires that may be caused by prior desires, reasoning, or deliberation. He then argues that motivated desires are desires only in the vacuous sense, since whatever explains them also explains the actions they purportedly cause. Among the factors that may explain them, he claims, are certain impartial rational principles expressive of one's self-conception as one temporally extended agent among many, i.e. principles of prudence and altruism. However, Nagel does not explain how a self-conception, or the principles that express it, can occurrently cause one to do something. In the absence of some such recognizably causal factor, the Humean is free to retort that since we are not rationally required to accept this self-conception, the principles that express it can be motivationally effective only if one desires in the *un*motivated sense to accept it. Then not only have these rational principles not been shown to be motivationally effective independent of desire; they have not been given a nonarbitrary rational justification, either. Nagel's project does contain the resources for an identifiable alternative to the Humean model, however: What Nagel could and should have said was that the description of a motivated desire may denote a particular episode of reasoning as a motivationally

[1](Oxford: Clarendon Press, 1975). Henceforth all page references to this work will be parenthecized in the text.

effective mental event whose causal influence depends on its intentional content. I do say this at length in Volume II, Chapter V.3.1.

Section 1 sketches the dilemma for a self-described Kantian who believes in the rational necessity of moral principle as a motive to action on the one hand, and in the universality but subjective contingency of desire on the other. I contrast Nagel's with Kant's solution to this dilemma: Both ground moral principle in a rationally inescapable self-conception that thereby motivates us to act on it. I evaluate the viability for this role of the particular self-conception that each proposes. Section 2 situates Nagel's analysis of prudential reasons for action in the context of his commitment to transpersonal rationality. Through analysis of his distinctions between timeless versus dated reasons on the one hand, and tenseless versus tensed judgments on the other, it traces his argument for the rational and therefore motivational inescapability of prudential reasons independent of any present or intermediary desires. Section 3 looks at Nagel's extension of his analysis of prudence to the analysis of altruism, the case in which the person on whose behalf one undertakes action is remote in space rather than in time. It examines his distinction between objective and subjective reasons, and evaluates the thesis that objective self-interested reasons give one reason to act on behalf of the person whose interests they denote. Nagel offers a second distinction, between the impersonal perspective on ourselves as one individual among many; and the personal, perspectival vantage point to which my analysis of funnel vision in Chapter II is often indebted. Here I examine Nagel's analysis of altruism as presupposing necessary connections among objectivity, impersonality, and universality; and his corresponding rejection of solipsism. I consider whether this argument yields the rational inescapability of ethical principle on which his thesis rests; and conclude that it does, at least, provide the conceptual resources for such an inference.

1. Nagel versus Kant

1.1. The Kantian Dilemma

The basic aim of Nagel's discussion is to show that to be rational is, among other things, to be capable of being motivated *directly* by altruistic principles and considerations – not merely by a desire of which these principles are the object. This aim expresses the transpersonal Kantian assumption that reason can be motivationally effective; not that desire is not, but merely that desire is not the only, and certainly not the most important, motivationally effective element in the self.

Nagel's idea is that if we are rational, altruistic principles and considerations themselves can inspire us to act on them; indeed, that this is part of what it means to be rational. By "altruistic" Nagel does not mean only acts of extraordinary heroism or self-sacrifice, but "any behavior motivated

merely by the belief that someone else will benefit or avoid harm by it." (16, n.1; cf. 79) These also include any act of "mundane considerateness which costs us nothing, and involves neither self-sacrifice nor nobility – as when we tell someone he has a flat tire, or a wasp on his hamburger." To demonstrate that altruism is a condition of rationality, Nagel tries to demonstrate that principles that prescribe such acts are as rationally "inescapable" as the laws of logic. He thereby means to show that moral principles themselves, broadly construed, rationally necessitate action; and therefore are motivationally effective.

I agree with Nagel's aim. But there are other ways of achieving it. My strategy, in Volume II of this discussion, is to analyze, not the inescapability of the principles, but rather their centrality in defining and exemplifying what transpersonal rationality is. Whether transpersonal rationality itself is inescapable is a moot question (but I doubt it).[2]

Nagel's way of approaching the problem of moral motivation encounters a dilemma almost immediately. He says,

> It may be thought that this excludes from an essential role in the foundation of ethics the factor of desire (although it is a mystery how one could account for the motivational source of ethical action without referring to desires). The problem about appealing ultimately to human desires is that this appears to exclude rational criticism of ethical motivation at the most fundamental level. As ordinarily conceived, any desire, even if it is in fact universal, is nevertheless merely an affection (not susceptible to rational assessment) to which one is either subject or not. If that is so, then moral considerations whose persuasiveness depends on desires depend ultimately on attitudes which we are not required to accept. On the other hand, the picture of human motivational structure as a system of given desires connected in certain ways with action is a very appealing one, and it can seem that any persuasive justification of ethical conduct must find its foothold in such a system (5).

In this passage Nagel's ambivalence towards the belief-desire model of motivation is evident. On the one hand, he literally cannot imagine how we could explain action without reference to desires; on the other, he accepts the Humean view of desires as too subjective and contingent to be subject to rational assessment and criticism. Yet the prevailing conception of motivation as a system of desires causally connected with action is plausible and appealing.

The dilemma for Nagel is that moral motivation must be as rationally inescapable as the truths of logic, if moral principles are to have the same stringency as rationality in general. By contrast, the Humean conception treats

[2]Thus my argument does not conflict with the important work of Kahneman, Tversky, et. al., that shows that actual agents do not reason decision-theoretically.

desires as universal on the one hand, but as idiosyncratic, arbitrary, contingent, and impervious to rational assessment on the other. Because desires, on the belief-desire model, are universal, any account of moral motivation must refer to them. But this implies that any account of moral motivation as rationally inescapable must therefore refer to a type of motivation that is not subject to rational assessment at all.

Nagel's dilemma is one with which any Kantian rationalist can sympathize. On the one hand, moral principles need to be universally applicable, objectively valid, and logically or conceptually necessary (or at least "inescapable," to use Nagel's term), if they are to do the necessary work of coordinating the behavior of different individuals and resolving conflicts among their disparate interests. After all, one wants to be able rationally to require of moral agents not only that they behave rightly, but also that they behave reliably. On the other hand, one wants an internalist account of reasons that will explain how these principles can causally effect action. These two desiderata seem *prima facie* incompatible. Although it is not inconceivable that some set of principles might satisfy the conditions of universality, objectivity, and necessity, our adherence to them need satisfy none of these conditions. Then how can we resolve the universality, objectivity and necessity of moral principle with the sporadic and seemingly nonrational motivation of particular agents to act on it?

Simply declaring that moral principle provides one with a reason for action is not enough, if the reason is not one the agent has in mind, or is one that has no compelling force for her even if she does. The externalist agrees that moral principle provides one with a reason for action; he denies merely that this reason in fact must motivate the agent to act. Even stipulating the agent's cognizings of the principle seems insufficient: the externalist can either deny that cognizings are motivationally effective mental events; or, even if they are, that they are any less arbitrary, contingent, transient, or idiosyncratic than desires. If cognizings and desires are similar in this regard, then in their absence, it is hard to see how we can be obligated to act on a principle of which we are, at least for that moment or circumstance, unaware, or by which we are uninspired, no matter how universal and necessary it is.

Kant's solution to the dilemma was to show consistent adherence to universal, objective and necessary moral principle to be a necessary condition of transpersonal rationality. He argued that agents who fail the requirement of consistent adherence could be shown to be defective in reason as well:

> [I]f reason solely by itself is not sufficient to determine the will; if the will is still subordinated to subjective conditions (certain drives) which do not always agree with the objective ones; if, in a word, the will is not *in itself* completely in accord with reason (as is really the case for human beings); then actions which are recognized to be objectively necessary are

> subjectively contingent, and the determination of such a will in accordance with objective laws is *necessitation*.[3]

Here Kant acknowledges the force of the externalist's claim, that reasons may have objective validity without our being moved to act on them; but also suggests that our *recognition* of their objective rational validity may compel or necessitate us to act on them despite our resistance. When our actions fail to accord with what reason requires, it is because "certain drives" interfere. We will see that something like this line of reasoning is attractive to Nagel, too.

Nagel begins by describing two possible solutions to the dilemma. The first is to reject and replace the belief-desire model of motivation. That is the solution I choose in Volume II of this discussion. The second is to retain the belief-desire model of motivation, but find a basis for distinguishing those desires that are susceptible to rational assessment from those which are merely arbitrary inclinations. This is the alternative Nagel chooses. He describes his task as follows:

> I shall propose that the basis of ethics in human motivation is something other than desire; but this factor will itself enable us to criticize certain desires as contrary to practical reason (5).

It is important to get clear about what Nagel is and is not saying in this passage. He is not saying that moral motivation is something other than desire. He is saying that the *basis* of moral motivation is something other than desire. This basis, whatever it turns out to be, can then be used to criticize desires as contrary to or in conformity with reason. But the first clause does not commit Nagel to repudiating the belief-desire model of motivation as he renders it ((a), 5), and the second clause suggests that he embraces it. Thus it seems that this basis of ethics will not provide an alternative *motivational* model for moral action. It will instead propose a new *criterion* for distinguishing those desires which are rational from those which are not. This introductory statement of Nagel's project does not portend a rejection of the Humean model, but rather an improvement on it.

What does Nagel mean by a "basis of ethics" such that it might have such a critical and rational role in evaluating desires? He characterizes it as a set of "motivational requirements on which to base ethical requirements" (5); and as "susceptible to metaphysical investigation", as carrying "some kind of necessity"(6). He also insists that the hold of these motivational requirements on us must be deep, and essentially tied to the ethical principles themselves

---

[3]Immanuel Kant, *Grundlegung zur Metaphysik der Sitten* Herausg. von Karl Vorländer (Hamburg: Felix Meiner Verlag, 1965) Ac. 412-413. Since I now comment on what I believe Kant actually to have said, rather than – as in Chapters I and V – on what other philosophers have gleaned from (to my mind faulty) translations of Kant's writings, I now work directly from the German original and offer my own translations. Henceforth page references to the Academy Edition of this work are parenthecized in the text, preceded by "G".

and to the conditions of their truth (6). The basis of ethics that serves as a standard for criticizing desires, then, consists in a set of independent requirements on desires which are metaphysical and in some sense necessary, and intrinsically connected to the ethical principles they ground. They also impose certain requirements that ethical motivation must meet.

However, the basis of ethics Nagel has in mind is not a justification of altruistic principles. For justification implies persuasion, and this depends on particular, arbitrary and idiosyncratic influences that get people to change their minds. Nagel, by contrast, means to show the inescapability of these principles *regardless* of the particular empirical influences at work. He means to furnish a psychologically pervasive foundation for ethics that demonstrates the rational pervasiveness of the principles it engenders. Thus he rejects any account of moral motivation that requires a prior, externalist motivational influence independent of moral principle itself. The foundation he seeks will explain the motivational influence of moral principles on action by putting those "principles themselves at the absolute source of our moral conduct." (11)

Nagel suggests that the rational inescapability of a moral principle can be shown by our inability to reject it once we become aware of it. But there are several arbitrary desires I might list that I would be equally unable to reject once I became aware of them (so I won't call them to mind by listing them). The rational necessity, or inescapability, of moral principle requires more than this. It requires, not only our inability to reject it once we become aware of it, nor even, in addition, that this inability be explained by our cognitive grasp of its content; but also that this inability – and corresponding action in accordance with it – be explained by our recognition of the *rationality* of its content *per se*. In this case, Nagel faces the daunting task of explaining how an abstract object of thought – the property of being recognizably rational – can causally influence anything at all; and, supposing it can, why it should influence an agent's consciousness and action at one particular time and place rather than some other.

1.2. Two Self-Conceptions

Nagel must, then, defend a foundation for ethics that has the following four features. First, it does not replace desire as the primary motive of action. Rather, second, it grounds and evaluates desires with an eye to their rational conformity to normative moral principle. Third, it *thereby* exerts some motivational influence on action. And fourth, it satisfies the rationality requirements of necessity, ethical connectedness, etc. listed above.

Nagel's chosen model for providing such an account is Kant's ethics. Kant is an internalist without being a Humean, because he insists both that moral principle is motivationally effective, and also that motivationally effective moral principle does not presuppose any desire of which it must be an object. Kant's idea is that moral principles express the agent's self-conception as free.

They are motivationally effective because we accept them, and we accept them because they express a self-conception with which we identify. By accepting a certain self-conception, we accept the corresponding moral obligations that follow from it, and are motivated by them to act accordingly.

Nagel assigns the same, psychologically central role to a particular self-conception in order to explain the motivational influence of the principles derived from it. Just as, in Kant's view, we are inescapably committed to our self-conception as free, similarly in Nagel's view, we are inescapably committed to our self-conception as merely one person among many, equally real ones. This self-conception explains the occurrence of altruistic behavior just as, in Kant's view, our self-conception as free agents explains our governance by the categorical imperative. In both views, the existence of genuinely moral conduct is presupposed rather than called into question. Therefore, an explanation of its existence, rather than a justification of its possibility, is what is required.

It is important to note that Nagel appeals to a self-conception, rather than to a conception of the self as those notions were distinguished in the General Introduction, above, to explain the fact of moral or altruistic conduct. That is, Nagel cites a pervasive way in which we think of ourselves, rather than a theory of what we are like in fact, to explain why we sometimes act to benefit others. The difference is important. The validity of a conception of the self is determined by its capacity to explain the psychological facts, generate testable hypotheses, and accrue confirmation from the results. Thus it gains or loses support depending on the plausibility of the link it proposes between what human agents are and what they do; and its truth or falsity is independent of what any particular individual thinks about it. So, for example, if the true conception of the self is a Humean one, according to which action is always motivated by desire-satisfaction, then at least a certain *kind* of altruistic action such as that performed by some whistleblowers discussed in Chapter VI.5.2 – i.e. that which involves no satisfaction at all to the agent who performs it, is impossible in theory. On the other hand, if a Kantian conception of the self is the correct one, then we may receive with skepticism Abraham Lincoln's protest that he was only saving the piglets from drowning in the mud in order selfishly to insure his peace of mind.

By contrast, the role Nagel assigns to a self-conception in his scheme is dependent on how individuals think of themselves, i.e. on whether or not they hold a certain belief. If they believe of themselves that they are no more or less real than other human agents, then they may invoke this belief to evaluate the integrity of their desires in relation to it. If those desires diverge from the agent's self-conception as one among many equally real agents – if, for example, one desires to use others in ways that overlook their capacity for emotions and thoughts as complex as one's own, then they express a belief that the agent is more real than others. This, for Nagel, is practical solipsism,

and it represents an internal dissociation from the impersonal standpoint expressed by that self-conception. Nagel will argue later that such an internal bifurcation in the self is irrational, contrary to practical reason.

But how rationally inescapable can this particular belief be? Compare the analogous role of Kant's proposed self-conception with Nagel's. According to Kant,

> it is impossible to conceive a reason with its own consciousness in relation to its judgments that receives guidance from outside, since then the subject would ascribe the determination of its power of judgment not to its reason, but instead to an impulsion. Reason must view itself as author of its principles, independently of foreign influences; consequently, as practical reason or the will of a rational being, it must be viewed by itself as free. (G, Ac. 448)

Kant's idea is that conceiving ourselves as self-determining is inherent in the nature of reason itself. To reflect or make a decision rationally is, according to our prereflective conception of these activities, to act voluntarily, without external coercion. In fact there is no behavior identifiable as an action we can perform that we do not experience as performed freely in this sense. So *all* deliberate actions, not just altruistic ones, are performed under the presumption of freedom. Moreover, they are performed under the presumption of freedom *because that is the way reason operates*. To experience ourselves as coerced or compelled is automatically to ascribe the source of causal determination to something other than reason, something external and nonrational. So for Kant, our self-conception as free is both universal, in that it governs any deliberate action whatsoever, and also inextricably connected to our rationality, in that it is a consequence of the proper functioning of reason itself.

By contrast, Nagel's proposed self-conception as one among many equally real individuals cannot be supposed to govern *all* actions, just altruistic ones. My own preoccupations may well take a back seat while I am being moved to warn you about the wasp in your hamburger; but while I am doing my tax returns they surely do not. By itself this is unproblematic. Similarly, the fact that *modus ponens* does not govern all forms of inference does not undermine its rational necessity. However, it is not obvious that Nagel's proposed self-conception is inherently connected to the functioning of reason in the way that Kant claims for our self-conception as free. Empiricism adduces rational arguments that Nagel's proposed self-conception is unfounded, and solipsism adduces further rational arguments that it is false. Moreover, it has been argued convincingly that the arguments for solipsism

are, strictly speaking, irrefutable.[4] If solipsism is rationally justified, as Descartes suggests, then the solipsist's dissociation from the impersonal standpoint may be rationally justified as well. So the inherent connection of this proposed self-conception with rationality must be demonstrated rather than merely asserted.

Furthermore, it is not clear that we are inescapably committed to this self-conception. Pathological narcissism is a pervasive psychological malaise in which precisely what is lacking is a conception of other people – their needs, feelings, and interests – as just as real as one's own. I argue in Chapter VIII.3.2.4 below that a narcissist may accept a self-conception such as the one Nagel proposes, without any vivid sense of the reality of the actual people she is thus conceiving. Nor are we always cognizant of this self-conception under relevant conditions, even supposing we are in some sense committed to it: It may be more vivid when my self-esteem is low or I am feeling humble, and weaker when I am preoccupied with my own interests or feeling self-important. All our experiences are bounded by the requirements of logical consistency, but not all our actions are governed by our self-conception as one individual among many equally real ones. Hence Nagel needs to explain the sense in which we are inescapably committed to this self-conception. Otherwise this commitment itself will be as vulnerable to the charges of contingency, idiosyncracy, and transience as the Humean desires it is intended to replace.

Now Nagel might reply that there is a difference between psychological inescapability and rational inescapability. That we very often fail to reason according to the rules of rational inference does not undermine the rational inescapability of *modus ponens*. Similarly, he might say, the fact that we often fail to conceive others as just as real as we are does not undermine the rational inescapability of doing so. If we are being rational, then, he might claim (following Kant's strategy), we must conceive others in this way. This would make the task of demonstrating the connection of this self-conception to rationality all the more pressing.

Finally, suppose we are inescapably committed to this self-conception. Does it inherently connect with altruistic behavior in the way Nagel claims? Nagel's compelling insight is into the connections among a certain kind of unselfishness, objectivity, and impersonality. My experience of the reality of another's interests both overrides, for the moment, my preoccupation with my own, and also, thereby, enables the interests that are a reason for him to act to become a reason for me to act on his behalf. Viewed from the impersonal standpoint, these interests are equally compelling regardless of which one of

[4]John Wisdom, "Philosophy and Psychoanalysis, " in *Philosophy and Psychoanalysis* (Los Angeles: University of California, 1969), 169-180; also *Other Minds* (Los Angeles: University of California, 1965).

us I happen to be. These interests acquire objective validity – and thereby, rational inescapability – because their status as a reason for me to act is not diminished by the priority I subjectively accord to my own. It is the combination of these three elements that rationally compel me to action: At the moment I spy the wasp alighting on your hamburger, my own concerns recede into the background. The thought that you may bite down on it occupies my attention totally, and the necessity I feel of preventing you from doing so feels overwhelming and unquestionable. According to Nagel, this is because I impersonally recognize your interests as just as real, vivid and worthy of promotion as my own.

But may it not also happen, when I am feeling vulnerable and assaulted by the reality of my situation, that this vivid reality impels me to wince and withdraw into solitude or privacy? In fact, does it not often happen that we feel assaulted, overloaded, overwhelmed and even exploited and manipulated by vivid awareness of another's need, misfortune, or imminent danger? In such cases, my self-conception as one among many equally real individuals may cause me to feel invaded, and to restrict or muffle rather than respond to the reality of the other. The worry then surfaces that despite the rational content of this self-conception, my actions may depend on other psychological events only contingently connected with it that may diminish or subvert its motivational influence on me. In order to show that altruism is a rational requirement on action, Nagel eventually will need to put all of these worries to rest.

2. Prudence

2.1. Transpersonal Rationality and Action

Altruistic reasons, as Nagel points out, are parasitic on self-interested ones: In order for me to act to benefit your interests, you must already have such interests. So, Nagel reasons, the form of altruistic reasons for action will depend on the form of prudential ones. Therefore, an answer to the prudential question of whether our future interests provide us with present reasons to secure them, should precede the answer to the altruistic question of whether others' interests provide us with our own reason to secure them. Nagel's first task, then, will be to show that our own future interests give us rational motives for present action to secure them, without any temporally intermediate or present desires interposed between them.

But the metaphysical analogy Nagel proposes between altruism and prudence has even broader ramifications than this. In defending both altruism and prudence, Nagel implicitly rejects two defining elements in the Humean conception of the self. The first has to do with my spatial relation, as a bounded three-dimensional subject, to other discrete subjects who inhabit the same space. To whose interests should I give priority? Does the spatial

proximity and intimacy of my own interests and desires give me reason to accord them motivational priority as well, as I argue in Chapter XIV that Hume himself suggests? By defending the possibility of altruism, Nagel rejects the thesis that each human subject necessarily views her own interests as overriding in importance the interests of spatially discrete and relatively remote others.

The second defining element in the Humean conception of the self has to do with my temporal relation, as a bounded temporal subject, to past and future subjects who successively inhabit the same continuous time-line. Again, to whose interests should I give priority? Should I satisfy my present desires simply because of their temporal proximity to me? By defending the possibility of prudence, Nagel rejects the doctrine of pure time preference, that I should view satisfaction of my present desires as overriding in importance, because of their temporal proximity, the satisfaction of my future desires. Thus Nagel proposes to defend a conception of human action as guided by a transcendent, impersonal perspective on its own spatiotemporal limitations; that is, by transpersonal rationality.

It is possible to understand this transpersonal, spatiotemporally transcendent perspective as the very embodiment of rationality, according to one early analysis of what rationality is. According to Jonathan Bennett,[5] this perspective is what distinguishes human agents as rational from other animals who appear to execute meaningful sequences of intentional actions that promote their common, long-term welfare, as bees do. Bennett argues that bees' behavior is biologically programmed stimulus-response behavior, independently of a genuine ability to conceive intentionally the elaborate sequence of plans that their behavior in fact carries out. By contrast, what makes us rational is our ability to make dated and universal judgments; to conceive any such act as a spatiotemporally localized instance of an abstract type of action which itself is not restricted to the spatiotemporal location of any particular token. This enables us, first of all, to range in thought over all such possible tokens, backward and forward in time and space; i.e. to connect conceptually what occurs in the indexical present to a possible or actual past and future, and to spatial locations other than this one. Second, it thereby enables us to abstract any such event or state of affairs from any particular spatiotemporal location at all, i.e. to understand it as a genuine and consistent abstract concept and apply it back to concrete circumstances accordingly. To do this, Bennett argues, is to exercise our capacity for theoretical reason:

> This is what generalizing and talking about the past have in common: they are both departures from that which is present and particular. This common feature is what links them with rationality. The idea of rationality is that of the ability, given certain present and particular data,

[5]*Rationality* (London: Routledge and Kegan Paul Ltd., 1964).

> to unite or relate them with other data in certain appropriate ways. This is the Kantian idea of concepts as unifiers, binders-together, creators of a *multum in parvo*. For there to be a 'multum' one must at one time intellectually possess more particular data than are present to one at that time, and for it to be 'in parvo' one must have rules or universal statements under which the particular data of which one is possessed can be subsumed. Thus: dated judgments and universal judgments.[6]

Thus Nagel's defense of prudence and altruism can be understood as an attempt to exhibit and defend the connection between our ability to think abstractly and impersonally, and our ability to act, against two views – externalism and Humeanism – that deny that there is any such connection. The great challenge for such a project is to articulate this connection in such a way as to clarify how the abstract objects that we think about or believe – propositions, principles, concepts, reasons – might enter into a motivational explanation of concrete and particular actions.

2.2. Nagel's Version of the Belief-Desire Model of Motivation

The anti-prudential and anti-altruistic view Nagel targets is recognizable as the belief-desire model of motivation. Nagel describes this model as asserting that

> all motivation has desire at its source. The natural position to be opposed is this: since all motivated action must result from the operation of some motivating factor within the agent, and since belief cannot by itself produce action, it follows that a desire of the agent must always be operative if the action is to be genuinely his. Anything else, any external factor or belief adduced in explanation of the action, must on this view be connected with it through some desire which the agent has at the time, a desire which can take the action or its goal as object. So any apparently prudential or altruistic act must be explained by the connection between its goal – the agent's future interest or the interest of another – and a desire which activates him now. Essentially this view denies the possibility of motivational action at a distance, whether over time or between persons. It bridges any apparent gaps with desires of the agent, which are thought to supply the necessary links to the future and to external situations. (27-8)

The view Nagel means to criticize, then, consists in the following, Humean line of reasoning:

> (1) All action must be caused by some (present) motivating event;
> (2) belief by itself cannot cause action;
> (3) therefore desire must cause action, and

[6] *Ibid*., 85.

(4) any other factor that may explain the action must be connected with a desire the agent has at the time of the action to perform that action (or achieve its goal).

(5) Therefore, any prudential or altruistic act must be explained by connecting its goal with a present desire of the agent.

Nagel depicts the belief-desire model as assuming that either a present desire must motivate action ((3)), or else there can be no present event to motivate action at all, since belief is impotent to do so ((2)). The alternative to a present motivating desire, as he sees it, is "motivational action at a distance, over time or between persons."

Now this alternative needs to be scrutinized very carefully. At first glance, it appears that Nagel is proposing to explain human action in terms of a kind of causation that is highly controversial regardless of the type of phenomena to be explained. That is, it appears that by "motivational action at a distance," Nagel means to refer to a species of causation at a distance: a remote mental event that has a proximate causal influence on my action with no intervening causal variables. In the case of altruism, the remote cause would be someone else's interest or occurrent desire as a proximate causal influence on my action. But in the case of prudence, the remote cause seems even more implausible. It would have to be my own future interest or desire as a future cause of my present action. This would be to advocate reverse causation. Call this the *implausible scenario*.

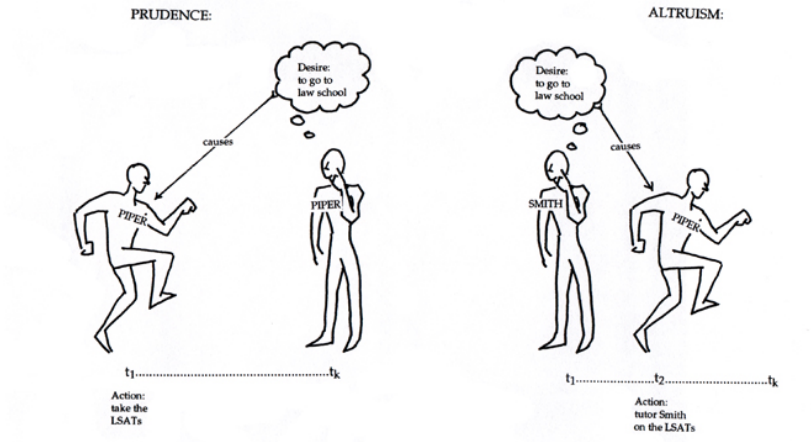


Figure 6. The Implausible Scenario

This way of describing both cases raises the same questions you might raise if I were to tell you that a certain book had strongly influenced my actions,

without clearly indicating that I had actually read the book: What exactly, you might naturally ask, was my mode of access to its influential content? Did I in fact read it? Or did someone throw it at my head, effecting in me a brain concussion that explains my newly gentle and considerate behavior?

Similarly with the stipulated influence of someone else's – or my future – desire as a cause of my present action. The question naturally arises: Of what sort might my access to this spatially or temporally remote desire be? In the altruistic case, did you communicate your desire to me in a linguistic utterance, or through body language? Or does your desire have some special power to move me independent of these mundane ways of getting people to do things for us? In the prudential case, am I simply expecting or predicting my own future desire? Or do I have some special ability to foresee my future? Or do my future desires have a mysterious power to reach back through time and affect my present behavior? In either case, how likely is it that any such remote desire might influence my action, without my having any of the usual modes of epistemological access to it at all?

The fact of altruism is not adequately explained by hypothesizing a causal connection between your desire and my action, without at least one intervening causal factor, namely my apprehension of your desire – i.e. my occurrent, true belief that you have it. Similarly, the fact of prudence is not adequately explained by hypothesizing a causal connection between my future desire and my present action without at least one intervening causal factor, namely my occurrent, true belief that I will have it. Nor is it feasible to concede the necessity of such an intervening factor while denying it causal efficacy by fiat. In any *plausible scenario*, my present apprehension of the relevant desire is a necessary link in the causal chain connecting that desire with my present action directed to its satisfaction:

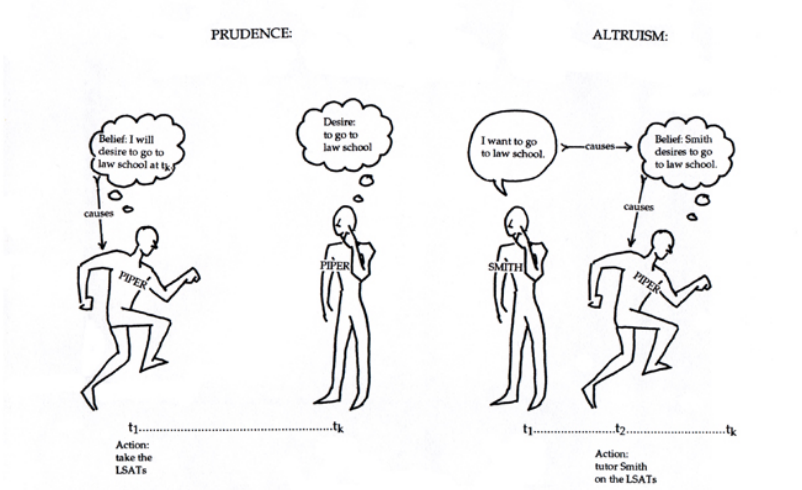


Figure 7. The Plausible Scenario

If Nagel were to grant this much, the question then would become whether it was the intentional content of the apprehension, or the event of its occurrence, or both, that was doing the causal work. If it were the mere event of its occurrence that had motivational influence, then it would be subject to the same contingent and idiosyncratic causal influences as any other mental event. On the other hand, if it were the intentional content that was supposed to be motivationally effective, then rational considerations should motivate action just in case I correctly apprehended that intentional content *as rational*, and not otherwise.

For if I do not apprehend an intension as rational when in fact it is, then either I will be motivated to act on it for irrelevant reasons, or else there is no guarantee that I will be motivated to act on it when and whenever I apprehend it. And if I do apprehend it as rational when in fact it is not, then I may be motivated to act on it when and whenever I apprehend it, but this regularity itself will be causal without being rational. In each of these cases, moral motivation will depend on contingent cognitive influences independent of the rational inescapability of the intension itself; and it will then be the contingent event of my apprehension of the intension as rational, not the rationality of the intension itself, that is doing the causal work. Then moral motivation will depend on the unpredictable occurrence of this contingent mental event, i.e. of apprehension, and will be no more rationally inescapable than before. So, more specifically, in order for the rational content of a desire to be doing the requisite causal work, I must be motivated to act on it when

and only when I apprehend it as rational; and I must, other things equal, apprehend it as rational when and only when it is.

But initially to grant the assumption, that my apprehension of your or my future desire is a necessary intervening link in the causal chain connecting that desire with the action motivated by it, is to reject part of Nagel's stated project. That is, it is to *agree* with the first premise of the belief-desire model that Nagel claims to attack, i.e. that all action must be caused by some present motivating event. Premise (1) does not stipulate that this event must be a desire; conceivably it might be an apprehension, recognition, belief, or perception instead.

We shall see, however, that in explicating his own view, Nagel himself sometimes uses such locutions that implicitly support premise (1), and sometimes explicitly rejects it. Although Nagel will distinguish clearly between the motivational conditions governing dated, present-tense judgments and tenseless judgments of what we have reason to do, he does, nevertheless, introduce such intervening causal factors at several points in the course of his own analysis of each. Hence he himself sometimes concurs with this premise of the belief-desire model. Nagel's analysis in fact constitutes an attack, not on premise (1), which he implicitly accepts for his own account (as he does premises (4) and (5)); but on premise (2), that belief by itself cannot motivate action, and on premise (3), that desire must. Although he claims to show that the structure of means-end reasoning itself can influence action, I propose that this is not where the true force of his arguments lay. Instead, they suggest that *certain kinds of occurrent belief – i.e. about what reason requires and about what kind of beings we are – can motivate action, namely those beliefs that have the status of reasons.* So it is in Nagel's implicit assumption that occurrent belief rather than desire might motivate action that his analysis does, in fact, suggest the outlines of a genuine alternative to the belief-desire model.

I say that Nagel's account contains the resources for such an alternative. But this is not the project he explicitly undertakes. He claims to attack premise (1), that all action must be caused by some (present) motivating event, and so to defend the possibility of literal "motivational action at a distance." This degree of divergence from the belief-desire model will be difficult to defend, for it suggests that not only are present desires unnecessary to motivate action; any other kind of present mental event is equally unnecessary. Thus it implies that our own future interests, or those of another, could cause us to act at the appropriate time and place, without our being in some sense occurrently aware of them at the time of action. I have described this scenario as implausible, and not only because Nagel offers no alternative mode of epistemological access to these interests. In addition, he offers no alternative factors to explain why one performs a particular action at a particular time and place on behalf of those interests, rather than a slightly different one half an hour later or two blocks away.

Nagel attends to each of the first three premises of the belief-desire model at different points in the text without clearly marking his intentions. He sometimes addresses premise (1), sometimes premise (2), sometimes premise (3), and sometimes neglects to distinguish among them. In working through this complex analysis, it will help to ask the following questions repeatedly: What present, causally effective mental event is in fact being invoked at any particular point to explain the agent's prudential or altruistic action? And if none is, what is being assumed to explain why it occurs precisely when and where it does?

2.3. Motivated versus Unmotivated Desires

Nagel argues that premise 2.2.(3) of the belief-desire model, that desire must cause action, is the consequence of a failure to distinguish between motivated and unmotivated desires. An *unmotivated desire* is one that "just assails us," such as an appetite or emotion. It has causal antecedents, as, for example, the causal antecedent of hunger is lack of food. But an unmotivated desire has no cognitive motivational antecedents, as, for example, a desire to go shopping may be preceded by a pang of hunger plus a belief that there is no food in the house. "The issue," Nagel observes, "is whether another desire always lies behind the motivated one, or whether sometimes the motivation of the initial desire involves no reference to another, unmotivated desire" (29). Nagel thinks it clear that not all actions need include unmotivated desires among their antecedents. But to establish this, he will need to establish an alternative foundation in the requirements of rationality for our temporally extended and impersonal self-conception as one person among many. He will have to show that this self-conception is not just another contingent and transient object of an unmotivated desire that assails us in moments of vulnerability or the need to escape the burden and intensity of our immediate and particular circumstances. Otherwise some other mental events – deliberating, reflecting, considering – that express this self-conception would be little more than the object of such an unmotivated desire – and so would come and go as they do.

Nagel defines a *motivated desire* as one caused by some prior psychological event. The event may be an antecedent unmotivated desire, such as hunger; or an antecedent motivated one, such as the desire for certain recipe ingredients; or it may be a decision or belief or judgment itself arrived at as the conclusion of a process of deliberation; or some combination thereof. Nagel's claim is that the explanation of a motivated desire is identical to the explanation of the action it causes. So, for example, my desire to take the LSATs is explained by my prior desire to go to law school, plus my belief that passing the LSATs is a necessary condition of admission to law school. My desire to take the LSATs is a motivated desire, in that what explains my actual

action of taking the LSATs is not my desire to take the LSATs, but rather my prior desire to go to law school plus my beliefs about how best to get there.

Nagel's point is that whenever we invoke a desire to perform act A in order to explain why someone performs act A, we have not really explained the performance of act A until we have an explanation for why the person desired to perform act A. And when we have that latter explanation, adverting to the desire to perform act A itself becomes irrelevant. Thus Nagel's claim constitutes an attack on the belief-desire model principle that we can explain a person's pursuit of a certain goal by ascribing to her a desire to pursue that goal. His argument is that it is trivially true, by definition of pursuing a goal, according to the belief-desire model, that the agent has a desire to pursue that goal; the presence of that desire is a conceptual truth. As we have seen in Chapter II, this is my complaint exactly about the belief-desire model. Therefore, Nagel goes on, in achieving a substantive understanding of why someone does something, we can effectively disregard her desire to do that thing as redundant, and concentrate on the antecedent conditions – the reflection, deliberation, beliefs, judgments, or prior motivated or unmotivated desires – that cause the action:

> That I have the appropriate desire simply *follows* from the fact that these considerations motivate me; if the likelihood that an act will promote my future happiness motivates me to perform it now, then it is appropriate to ascribe to me a desire for my own future happiness. But nothing follows about the role of the desire as a condition contributing to the motivational efficacy of those considerations. It is a necessary condition of their efficacy to be sure, but only a logically necessary condition. It is not necessary either as a contributing influence, or as a causal condition (29-30).

Now if motivated desires are *only* conceptual truths, then they have no motivational efficacy; and their antecedents are doing the only real causal work in precipitating action. Since unmotivated desires need not be present as causal antecedents of action, certain occurrent cognitive events – believings, considerings, recognizings or acceptings of principles or judgments – would seem, by a process of elimination, to suffice to motivate some actions. This possibility, if developed, would constitute an attack on both premise 2.2.(2) and premise 2.2.(3) of the belief-desire model. For by furnishing motivationally effective antecedent psychological events other than desire, it would not only challenge the thesis that desire must cause action (premise 2.2.(3)), but dispute the thesis that belief by itself cannot (premise 2.2.(2)).

However, even if motivated desires are not only conceptual truths but identifiable mental events as well, other such occurrent cognitive events would still have motivational efficacy as causal antecedents of action. For where unmotivated desires are absent, these motivated desires themselves would have to be the effect of prior acts of deliberation or evaluation. I show

in Chapter XI that this is precisely the status Richard Brandt assigns the concept of rational desire; and consider further the question whether this makes Brandt a Kantian or Nagel a Humean.

The textual evidence for the ontological status of motivated desires is ambiguous. The above passage denies that motivating desires play a necessary role in causing action. But shortly thereafter Nagel remarks that "[o]ften the desires which an agent *necessarily experiences* in acting will be motivated exactly as the action is. If the act is motivated by reasons stemming from certain external factors, and *the desire to perform it is motivated by those same reasons*, the desire obviously cannot be among the conditions for the presence of those reasons" (30; italics added). From this latter passage we can infer that the agent necessarily experiences motivated desires that really are motivated by prior psychological events. This means that they cannot be merely conceptual truths; they are substantive experiences as well. And later, in discussing the motivational role of present-tensed practical judgments that one has reason to act, Nagel says, "This judgment possesses motivational content, for one then regards the undertaking as justified, and this is sufficient to explain one's wanting it to happen, be happening, or have happened. *Such a desire will form* even if one does not know what time it is" (70-71; italics added). Again the implication is that a practical judgment can cause a motivated desire, considered as an occurrent mental state, to exist.

Similarly, when Nagel later discusses the implications of trying to apply a subjective principle from the impersonal perspective, he remarks that although one may then be able to identify those of the agent's acts justified by the subjective principle, "this is a mere classification without motivational content – without the acceptance of a justification *for wanting* anything" (122; italics added). Here Nagel equates motivational content with the acceptance of a justification for wanting something, from which we can infer that the justified want has motivational influence. This, too, makes it more than only a conceptual truth. From all four passages taken together, it would seem to follow that motivated desires do not necessarily play a causal role, but do necessarily play an experiential role. From this it would follow, in turn, that we necessarily experience certain mental events, namely motivated desires, that do not necessarily have causal impact on action. But how a consciously experienced mental event could fail to affect us causally in some way – if only to alter our brain chemistry slightly – remains obscure.

Nagel has averred more recently that "sometimes a motivated desire is a conscious mental state or event, even though its motivational force depends on the reasons behind it." He also characterizes it as a "propositional attitude, and therefore an intentional state."[7] These statements imply that a motivated desire can be (but is not necessarily) a mental event denoted by a sentence of

[7]Private communication of 20 September 1991.

the form, "I desire that P," where "P" is the description of my performing the action in question; and that this mental event has motivational force only if the practical judgments that cause it to form do. So a motivated desire *is* among the causal antecedents of action only if

(1) it is a mental event; *and either*
(2) it is preceded by a motivationally effective practical judgment *or*
(3) it is preceded by unmotivated desires.

We have already seen that unmotivated desires (2.3.(3)) are unnecessary as causal antecedents of action. Let us then consider 2.3.(2). Nagel claims that practical judgments that one has reason to do something always have motivational efficacy, so motivated desires that are mental events preceded by such judgments always do as well. If practical judgments that one has reason to act always have motivational content, then they have it *whether or not that judgment itself is well founded*. So it is the occurrence of the judgment rather than the recognition of its rational content as rational that motivationally influences action. It is then possible that the occurrent cognitive mental events that precede a motivated desire might consist in mental ruminations, associations, and imaginings that cause the desire but do not justify it; and that on the one hand do not just assail us, but on the other provide no reasons for the actions we take in response to them, either. So as yet we cannot regard that content as a rationally inescapable requirement on action.

So far no link has been established between the actual rationality of a judgment and its motivational efficacy. The conjunction of 2.3.(1) and 2.3.(2) leaves open the possibility that an irrational or nonrational judgment might cause an equally irrational or nonrational motivated desire, which in turn might issue an irrational or nonrational action. Judgments of this kind cannot provide criteria for the rational evaluation of desire, nor, therefore, for rational altruism in action.

Finally, consider 2.3.(1). A motivated desire may, according to 2.3.(1) alone, be a mental event that is in itself causally impotent. But it is also possible that it may be causally effective *because* preceded only by causally effective occurrent cognitive events such as believings, considerings, or deliberatings. So even if motivated desires are more than mere conceptual truths, Nagel's view implies that practical judgments, and the occurrent cognitive events they involve, may be the sole motivationally effective psychological antecedents of action in some cases.

Nagel has intimated something of this kind all along. For example, he earlier defined altruistic action as "any behavior *motivated merely by the belief* that someone else will benefit or avoid harm by it" (16, n. 1; italics added). He also characterized his own position as "one which ties the motivation to the cognitive content of ethical claims, [and] requires the postulation of

motivational influences which one cannot reject *once one becomes aware of them*" (8; italics added). Similarly, in his critique of G. E. Moore's internalism, he concluded that

> if one wishes to tie the requirement of motivation influence to the truth-conditions of moral claims, with the consequence that *if someone recognizes their grounds, he cannot but be affected accordingly*, then a stricter motivational connection will be required (9; italics added).

These are only a few of the many passages in which Nagel seems to acknowledge that conscious beliefs – or rather, believings – and recognizings may be occurrent psychological events that can motivate an agent to action, just as desires may, and even where no desires are present.

This is enough to put a serious dent in the belief-desire model of motivation, despite Nagel's more modestly reformist intentions. Nagel will have vanquished premises 2.2.(2) and 2.2.(3). He will have shown that desires need not cause action, and that other kinds of things besides unmotivated desires *can* cause action in those cases where motivated desires do not. This is not, after all, merely to furnish a rationally critical ethical *basis* for such desires. It is potentially to eliminate them altogether from the causal account, and replace them – at least in some explanations – with states that are not desires at all.

But what kinds of states? Other kinds of physical events? Or abstract objects of thought? The following passage does not resolve this question:

> If considerations of future happiness can motivate by themselves, then they can explain and render intelligible the desire for future happiness which is ascribable to anyone whom they do motivate (30).

Are "considerations" identical to occurrences of "considerings"? If so, it is not farfetched to entertain them as identifiable motivational influences on action, and thereby to repudiate premise 2.2.(2). Or by "considerations" does Nagel mean "the propositions or principles considered"? In this case, we really do need to know how Nagel explicates the causal relations between the rational content of such abstract principles and the actions they are presumed to motivate.

In order to see how Nagel resolves this question, we must find out more about what differentiates rational desires from nonrational ones. Identifying this criterion will enable us to decide whether it itself can be motivationally effective, or whether it merely enables us to cull the bases of altruistic action from those of self-interested action. Nagel seems to have the latter possibility in mind, when he draws an analogy between beliefs as the material for theoretical reasoning and reason itself as its inference structure, on the one hand, and desire as providing the material for practical reasoning and "something besides desire" that "explains how reasons function" on the other (31). He says, "This element accounts for many of the connections between reasons (including the reasons which stem from desires) and action. It also

explains *those general desires which embody our acceptance of the principles of practical reason*" (31; italics added).

There are at least two important points to be drawn from these passages. First, Nagel clearly thinks our acceptance of the principles of practical reason are embodied in desires of a certain sort. This may mean either that

> (4) we express our acceptance of these principles by having certain desires;

or that

> (5) our acceptance of these principles are motivated by such desires.

2.3.(5) would surrender the field to the Humean for all practical purposes, so to speak. But even 2.3.(4) concedes to the Humean entirely too much. For if what our acceptance of these principles comes to is just that we have certain sorts of desires, then by definition these desires must be unmotivated, not motivated desires. And then whatever action follows from thus accepting them cannot exclude unmotivated desires from being among its motivational influences. In this case, unmotivated desires – to reason, reflect, or deliberate – *are* necessary causal antecedents of those occurrent rational activities, and hence are no more or less rationally inescapable than those activities themselves. And it did seem that Nagel meant to resist this conclusion in the passage from pages 29 to 30, quoted earlier.

The second point to notice in these passages is Nagel's description of something that plays the same, formal role for desire that principles of theoretical reason play for belief. Nagel believes that a reason *in itself* must be able to motivate action, not because it contains some further, independent motivational factor (like a desire) among its conditions of application. Again, on the face of it, this is controversial at the very least. A reason, like a proposition or principle, is an abstract object, a "consideration" with rational content. Abstracts objects are not a species of causal event. So it is not easy to see how by themselves they might make any such event occur – or, in case they might, how they might bring it about that the actions they cause occur at just the spatiotemporal locations they do.

2.4. Means-End Reasoning and the Extraordinary Interpretation

Nagel proposes that the motivational efficacy of reasons derives from "the principle governing their derivation from" the "conditions for their existence" (32), i.e. the principle of means-end reasoning; and that "[a]ny acceptance of a reason for action must conform to the general principle concerning means and ends" (35). By this he means that certain ends of action provide reasons for undertaking the actions required to bring them about. For

example, if I have a long-term interest in becoming a lawyer, then I have a reason for passing the LSATs and so a reason for taking them. The principle that governs this inference says, roughly, that I have reason to take the means that will advance my ends. This principle expresses clearly the basic idea behind the utility-maximization model of rationality that I have already examined in Chapters III and IV.

But notice that, here as in Chapters III and IV, this principle of means-end reasoning is *structurally independent* of the belief-desire model of motivation, since forward-looking propositional pro-attitudes toward my ends are not the only motives this model of rationality can accommodate. Forward-looking propositional attitudes as such, containing no favorable or unfavorable evaluation of their objects – anticipation or resolution, for example, combined with backward-looking motives such as moral conviction, concern, or fear may also motivate means-end deliberation about which actions will best achieve the ends in question. I discuss this distinction between forward- and backward-looking motives at greater length in Volume II.

Nagel argues that my *recognition* of the application of this principle, i.e. that my action is a means to the achievement of my ends, makes my reason for passing the LSATs – that it will help me become a lawyer – motivationally effective. This is a "general principle concerning means and ends" (35), namely that "[r]easons are transmitted across the relation between ends and means, and that is also the commonest and simplest way that motivational influence is transmitted. No further desires are needed to explain this phenomenon" (33).[8]

By the claim that reasons are transmitted across the relation between ends and means, Nagel means, then, that the end or goal provides a reason for taking the means to achieve it. Thus the circumstance of having a particular end furnishes a reason for that action identifiable as a means to its achievement (and not for actions that are not thus identifiable). And by the claim that this is also the way motivational influence is transmitted, Nagel means that the end in question also – thereby – provides the agent with a motive for achieving it.

Now if we interpret this thesis in the *ordinary* way, Nagel is observing simply that the occurrent thought of the end I wish to achieve motivates me to achieve it. This would provide a genuine alternative to the belief-desire model because it would not require that I have a desire of any sort, whether motivated or unmotivated, to achieve that end. Instead I might be motivated antecedently by the voice of conscience, or recognition of moral obligation, or a sense of compassion, or a reflective resolution to formulate and then achieve the end in question. However, my thought of my end is in the present; only

[8]However, not any end can furnish a reason for action, according to Nagel. Ends that are merely objects of desire do not.

the prognosticated end itself – i.e. the object I represent in my thought – is in the future. By having denied the necessity of a presently motivating mental event, denied that beliefs (or thoughts) can motivate action, and committed himself to defending motivational action at a temporal distance, Nagel has committed himself to the following, *extraordinary interpretation* of the above thesis: that it is not my present thought of the end that now moves me to act, but rather *the future end itself* that transmits motivational influence back to me in my present state!

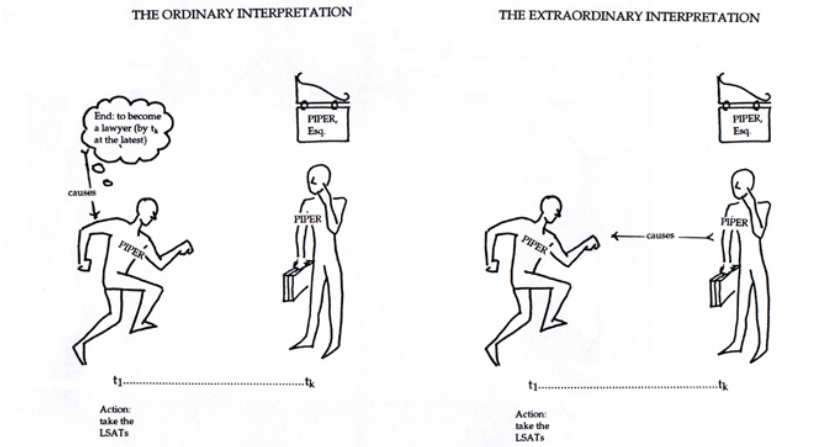


Figure 8. The Ordinary vs. the Extraordinary Interpretation

As frequently occurs in this complex text, Nagel offers us an exegetical choice at this juncture. First, we may refuse to take Nagel at his word, and so reject the extraordinary interpretation by turning our attention to turns of phrase that indicate that Nagel implicitly intends to advance the ordinary interpretation instead. For example, Nagel says, "The problem is how any considerations about the future, about the long-range outcomes of alternative courses of action, can affect an individual's behaviour in the present" (37). Again on the same page, he states, "The issue, by now a familiar one, is whether the effect on present action of beliefs about my future interests must be explained by an intervening desire, or whether the connection can be made through a requirement of practical reason by which actions are governed." In both of these passages – in addition to many others that could be cited, Nagel deploys phrases – "considerations about the future," "beliefs about my future interests" - that lend themselves to the ordinary interpretation, i.e. that a present mental event, namely an occurrent belief (or "considering") about my future, governed by requirements of practical reasoning, motivates me to act

so as to secure my future wellbeing. (Also see pages 36, 39, 40, 44, 45, *passim*, where Nagel uses the term "expectation" in the same sense.)

The problem with the ordinary interpretation already has been mentioned: To construe beliefs, considerings, or expectations as nothing but motivationally effective present mental events is to divest them of any necessary connection with rational rather than merely causal sequences of thought and action. In this case, prudential (and altruistic) action would be no more or less rational than motivation by unmotivated desires – unless Nagel can supply an account of how rational content motivates moral action with rational inescapability.

Second, we may opt for the extraordinary interpretation and see how far we can take it. The extraordinary interpretation requires us to make sense of the thesis that a future end – my securing of my wellbeing along some particular dimension – by itself affects me motivationally in the present. How could this possibly happen? Nagel's answer is that if an end is a genuine reason, then it is a *timeless reason*: It is a reason not just for one particular act-token, but instead for a class of act-types that promote that end; and that class contains no explicit temporal restrictions on when the act-token must be performed. Of course it must bear a certain kind of temporal relation to the end itself: the act-token cannot temporally succeed the end it is supposed to promote. But this relation itself is perfectly general, in that it does not restrict the occurrence of the act-token to any one temporal domain. Since a timeless reason applies to a temporally unrestricted domain of act-tokens, it cannot explain why a particular act-token that is justified by it is performed at the particular moment it is. By contrast, a *dated reason* is one derived from a timeless reason that obtains in the present at the time of action. Nagel also argues that dated reasons, in turn, imply timeless reasons plus a statement relating the reason to the time of utterance; and that to accept only dated but not timeless reasons demonstrates a failure to identify oneself as unified through time. Note that a dated reason as such does not entail a dated, occurrent event.

Thus Nagel explains the transmission of motivational influence from an end "backwards" in time to the action taken to secure it, by elevating the end in question to a general and unitemporal, or timeless, status. As Nagel expresses it,

> If there is a reason to do something on a particular occasion, it must be specifiable in general terms which allow that same reason to be present on different occasions, perhaps as a reason for doing other things. All such general specifications, whatever else may be true of them, will share a certain formal feature. They will never limit the application of the reason to acts of one sort only, but will always include other acts which promote those of the original kind. (35)

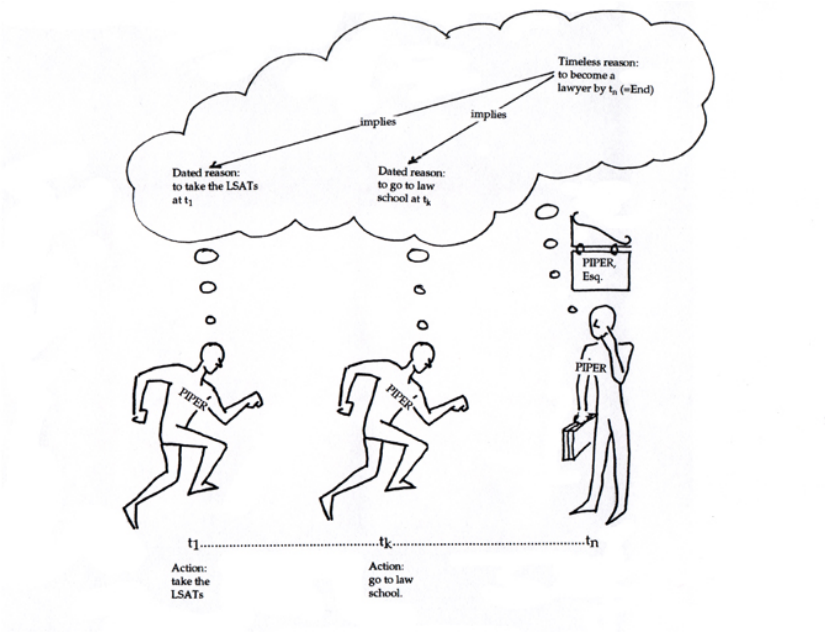


Figure 9. Timeless vs. Dated Reasons

So, for example, if becoming a lawyer is a reason for me to take the LSATs, it is also a reason for me to pass them, and also to go to law school. Moreover, it is a reason for me to work as a paralegal not just today, when I am feeling ambitious, but all summer, although the weather is pleasant and the outdoors beckons. Similarly (to take Nagel's own example), if my future vacation in Italy is a reason for me to learn Italian now, it is also a reason for me to speak it fluently when I am there. Finally, my future vacation in Italy, by virtue of being a reason for me to learn Italian now, becomes a reason for me to now enroll in Conversational Italian, Level I at the university.

Thus it is a consequence of what Nagel describes as the "timeless generality of reasons" that my future ends give me not only present, dated reasons to perform acts that promote them (e.g. to take the LSATs), but present reasons to perform acts (e.g. to work as a paralegal) for which I expect there to be a reason in the future. If my future end is indeed a reason, then it is always a reason; and *a fortiori*, it is a reason *now*.

But is this enough to defeat the unpalatable implications of the extraordinary interpretation? Even if Nagel is right about the timeless generality of reasons (and I think he is), conferring this status on certain ends is not sufficient to motivate one to act in their service without some further element, such as an occurrent thought, belief, or expectation about them, that

connects them to me in my present state and gives me epistemological access to them. Without some such element, Nagel is reduced to the externalist's claim that I have a reason for certain actions even if I do not know or believe that I do, and *a fortiori* even if I am not motivated to perform them. And indeed there are passages in the text that invite this reading. For example, when he says that "if the event is future this principle has the consequence that one has a present reason to promote it *simply* because there *will* be a reason for it to happen when it happens, and not because of any further condition which obtains *now*" (48), he implies that one has a present reason to promote this future event irrespective even of one's present beliefs about this event, and *a fortiori* irrespective of whether or not one knows one has a reason to promote it. This is another passage in which Nagel seems to mean to attack premises 2.2.(2) and 2.2.(3), under the guise of attacking premise 2.2.(1).

Nagel supplies one such an element by embedding my future interest in a self-conception as unified over time, which, he thinks, I have at any particular time. To view my future end as a timelessly general reason for present action, I must now believe I will be the same person then as I am now, and hence that my future interests are in some sense my present interests as well. Thus I gain epistemological access to my future interests and desires by embedding them in a self-conception I now believe holds true of me both now and later. He says,

> The hypothesis that all links to the future are made by present desires suggests that the agent at any specific time is insular, that he reaches outside himself to take an interest in his future as one may take an interest in the affairs of a distant country. The relation of a person to temporally distant states of his life must be closer than that. His concern about his own future does not require an antecedent desire or interest to explain it. There must already be a connection which renders the interest intelligible, and which depends not on his present condition but on the future's being part of his *life* (38-9; also see 42-3).

According to Nagel, it is because, from a standpoint of temporal neutrality, I identify with all past and future, equally real stages of myself as forming a single life that future, reason-providing interests provide timeless reasons for me now to act, even though those interests themselves are to be fulfilled in the future. So part of the ethical (strictly speaking, *meta*ethical) foundation Nagel proposes for evaluating the rationality of desire is the conception of oneself as unified through time, such that timeless reasons express values that define the self and in turn generate dated reasons to act at appropriate moments. If the motivationally effective desires one has at such a moment are justified by those reasons, then those desires are rational.

Is this enough to answer the questions raised about our relation to our future interests? In order for timeless reasons to explain why I now act on behalf of my future interests, Nagel's stipulation of an underlying self-

conception of myself as the same person later that I am now implies that I conceive myself as a person whose interests later are equally, therefore, my interests now. The first question would be whether this self-conception is accurate, i.e. whether I am in fact the same person later that I am now. But assume it is accurate. This does not necessarily imply that my interests later are the same as my interests now. Surely I can change my mind about at least some centrally important matters without renouncing my personal identity. Indeed, recall the argument from Chapter IV, that minimal psychological consistency requires only one long-term preference not subject to revision in ranking status on each occasion on which pairwise comparisons are made. But let us give Nagel the benefit of the doubt here as well.

Nagel does not claim, but the plausible reading of his argument suggests, that I now must know what those future interests are, in virtue of knowing my self-conception, in order for them to now motivate my present action. If my future interests are my present interests because my future self is as much part of my life as my present self, then my future interests are, at the very least, also *my present interests*. If they are also my present interests, then there is no great mystery as to why they should now motivate me to act to secure their future satisfaction, but also no reason to deny premise 2.2.(1), since this account is compatible with it. What motivates me to secure my future interests is the present desire to secure interests that are both present and future (or, on the broader, ordinary interpretation, perhaps the presently motivating thought of those interests as equally my interests now).

To ignore or repudiate this self-conception, Nagel argues, is to act only on those dated reasons that I acknowledge to obtain at the time of action – in which case I have neither reason to prepare for future eventualities, nor reason to regret those I have ignored, nor reason to repudiate later those I now know I will want to repudiate then (40, 42, 58). So even if such indifference to other stages of myself were possible, it would express a dissociation so radical as to subvert future-directed action altogether.

2.5. Tensed versus Tenseless Judgments

Nagel effectively concedes the necessity of a present mental event, alternative to desire, which can serve to motivate my present action, by introducing the distinction between tensed and tenseless judgment. He argues that a temporally dissociated (or pure time-preferential) self makes only present-tensed judgments about what it now has reason to do. These reasons will refer to ends or interests the agent now wishes to promote. Hence tensed judgments embed dated reasons in present mental acts of judging what one now has reason to do. However, Nagel argues, any such tensed judgment implies a tenseless one made from a standpoint of temporal neutrality – the standpoint of a self unified over time. This judgment embeds a timeless reason that applies derivatively to the particular act-token that now promotes

the ends or interests in question (49, 54). The tenseless judgment itself occurs at the particular moment the tensed judgment is made, but it does not refer to that particular moment any more than to any other. Thus its derivative status unfits it for explaining why the act-token occurs precisely when it does. For that we need the tensed judgment that implies it. So the present mental event alternative to desire that motivates my present action is an occurrent act of judging that I now have reason to perform that action.

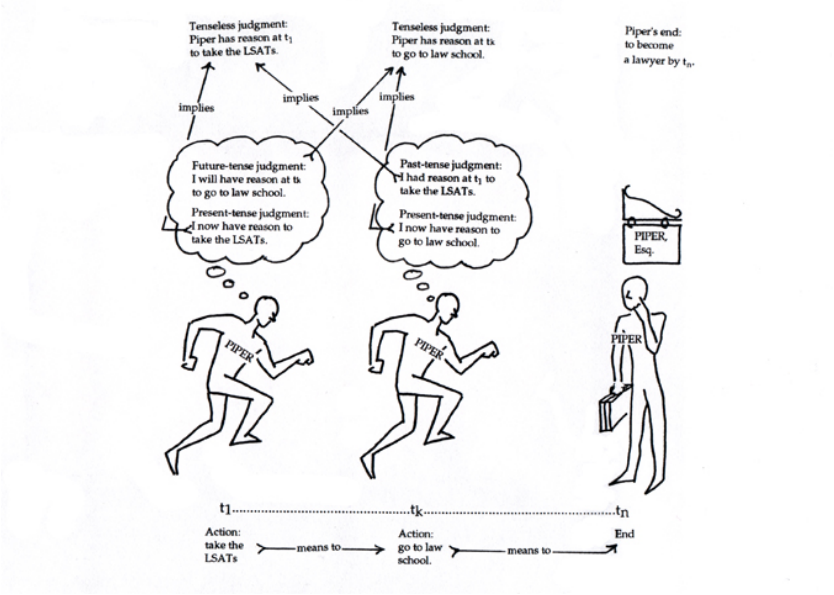


Figure 10. Tensed vs. Tenseless Judgments

Nagel considers the objection that a tenseless truth about a future time may not be true in the present, if the event to which the tenseless judgment refers is not one we now have reason to foresee:

> [I]t may be true on 15 May that there is an airline strike. But it may not yet be true on 1 May that on 15 May there *will* be an airline strike, and in that case how can it be even tenselessly true *on 1 May* that there is an airline strike on 15 May? And if that is not true on 1 May, how can it create a reason on 1 May to arrange alternative land transportation to Chicago for 15 May (54)?

Nagel considers two solutions to this question. The first would be the externalist's solution. It would be that in fact there is a tenseless truth about the future and a derivative, dated reason for acting in the present, but that the agent cannot know this. Just as the externalist would avow the existence of

reasons that do not motivate, he would similarly avow the existence of reasons the agent cannot know or recognize.

A strong reading of the externalist's solution would rely on the assumption of causal determination of the future, independently of the agent's ability to predict or know it. It would assume that there is a fact of the matter about whether there will be an airline strike on 15 May, independently of whether anyone, including the strikers, knows this; and therefore that reasons for arranging alternative land transportation to Chicago at a certain time themselves may exist timelessly, independently of when or whether anyone knows that the conditions they cite will obtain.

A weaker reading of the externalist's solution would say merely that, once the union decides to hold an airline strike on 15 May, there is a fact of the matter about whether there will be one, independently of whether the agent planning her trip knows this; and therefore, that once the union's decision is made, that her reason for arranging alternative transportation exists, independently of whether she in fact knows that the conditions it cites will obtain. Note that that Nagel's concept of a timeless reason strictly requires the first, stronger reading of the externalist's claim, because the weaker reading temporally localizes the reason to the actual existence of the conditions it cites. I do not regard the commitment to causal determinism inherent in the stronger reading as problematic. I mean only to point out that Nagel's concept of a timeless reason requires it.[9]

A second possible solution – the one Nagel accepts – is to say that the agent "has no reason to promote an end until it is true that the reason-predicate holds or will hold of that end" (54-55). This is in fact consistent with the first, externalist solution, because it observes the distinction between there being a reason for action and the agent's consciously having a reason for action. Nagel's solution proposes that a necessary precondition of consciously having a reason for action is that the agent's end provide her with that reason.

Now the strong externalist would say that an end that provides a *bona fide* reason for action is an end she has – timelessly, regardless of when she consciously adopts it. The weak externalist would say that she must

[9]Nagel later (62, n. 2) attempts to dispose of the objection that his view precludes the existence of future contingencies not determinable from present conditions, by proposing tenseless judgments of probability (or, as he puts it, the relative "definiteness" of a future event) for the existence of conditions that supply reasons for action. The problem is that the expected existence of such conditions must be assigned a certain degree of probability before they can become reasons for action; and that that assignment will change, according to the information about its probability available at the time. Hence judgments of probability are implicitly tensed judgments that become tenseless in proportion as *p* approaches 1. This presents a problem for Nagel's view only if it is implausible to assume the operation of causal determinism at the macro-level of action. I do not think it is.

consciously adopt that end before she can have it – and so before that end can be a reason for her action, whether or not she then recognizes it as such. But the internalist would reply that she cannot adopt that end – i.e. to get to Chicago despite the airline strike – until she knows about the airline strike on 15 May. Once she knows about the airline strike, she deliberately adopts the end of getting to Chicago despite it. But once she does that, she recognizes it as a reason to arrange alternative land transportation. So there is a reason, in this case, if and only if she has a reason; and she has a reason when and only when she consciously adopts it. So despite claims Nagel makes elsewhere (55), internalism, even of a not-obviously-Humean variety, does locate her reason for action within a particular temporal domain, i.e. between the time she learns of the pending airline strike and the time it occurs. Timeless reasons, on Nagel's account, are not timeless strictly speaking. Rather, they obtain within a specific temporal domain of a temporally restricted self, unified throughout that restricted time.

These two solutions thus reopen the Kantian dilemma, i.e. the conflict between the Kantian rationalist's requirement that reasons be universal, objective and necessary, and the internalist's that they be motivationally effective for a spatiotemporally localized and contingent agent. What Nagel needs here is a way of retaining the (relative) timelessness of the reason itself consistently with the agent's epistemically limited access to it. He should accept the strong externalist's solution and simply add that a reason itself can be (relatively) timeless even though the agent's epistemological access to it is dated. Nagel suggests but does not develop something like this possibility when he avers that

> although the conditions which confer a reason on an event have been conceded not to be timeless, *the value which that reason embodies is still timeless*, and can transmit its influence from future to past, once it is clear that the reason will be present (55; italics added).[10]

2.6. Motivational Content and the Extraordinary Interpretation

A tenseless statement expresses the sense of an assertion that may be made at different times and in different tenses; it expresses "what is asserted in common by past, present, and future [tensed] statements about the same circumstance or state of affairs" (61). It thereby expresses the standpoint of temporal neutrality that regards each stage of the speaker's life as equally real and equally a part of that person's life. We have seen that this conception of oneself as temporally extended then may be invoked as a criterion of adequacy for reasons for action:

[10] Elsewhere I argue that a distinction between the dictates of reason themselves as abstract objects and particular instances of reasoning in accordance with them is also part of Kant's solution to the question of how universal and necessary reason can cause particular act-tokens that conform to its moral prescriptions.

> I contend that one can ask of a practical principle or a reason for action whether it is consistent with the conception of oneself as a person extended in time, or whether the acceptance of it must be dissociated from that conception. Some principles and reasons are in accord with the conception, but others, I believe, are not, for their acceptance is not compatible with the neutrality of viewpoint toward different times whose possibility I have argued is essential. (63)

Accordingly, tenseless judgments made from the perspective of temporal neutrality that are reasons for action are judgments "that certain acts or desires on my part will be *justified* at that time." The judgment that

^certain acts and desires on my part will be justified at t_k^

is itself an example of such a judgment.

Nagel claims that a tenseless judgment "has motivational content. To accept a reason for doing something is to accept a reason for *doing* it, not merely for *believing* that one should do it" (64). This is the basis on which action and desire is to be rationally criticized (64-5). What is motivational content? The motivational content of a judgment involves a "commitment to act, or to desire" (64). In fact, Nagel says that a judgment that I have reason to do A itself includes the acceptance of a justification for doing it (65); and later defines motivational content as the acceptance of a justification for doing or wanting something (109). So a tenseless judgment that I have reason to do A itself includes, not only a justification for doing A, but also the *acceptance* of that justification for doing A; and this acceptance of a justification, Nagel argues, can be motivationally effective. Thus according to Nagel, tenseless practical judgments that I have reason to do A have causal efficacy built into them.

But is it true that tenselessly acknowledging that I have reason to do A can move me now to do A? Nagel first addresses the motivational content of present-tense judgments: judgments that I now have reason to perform some action. He argues that to deny the motivational content of present-tense judgments is to open the door to an infinite regress of justifications for doing what it is justified to *believe* one should do, and none of which then count as the conclusion of practical reasoning, namely that one should act accordingly. He acknowledges that motivational content does not necessarily imply motivational efficacy in every case, since "motivational interference" in the form of weakness of will, cowardice, laziness, etc. is always possible. But he goes further, by claiming that if no such interference is present and the agent still fails to perform the justified act, doubt is cast on the supposition that he actually accepts the judgment in the first place:

> It is undeniable that someone may acknowledge a reason for action and fail to act. Indefinitely many circumstances may explain this. Indeed, I

> believe that a practical judgment can sometimes fail to prompt action or desire without any explanation. Not every case of irrational behavior need be comprehensible. But in what sense can a judgment possess motivational content if its motivational efficacy can be blocked in indefinitely many ways (65)? ... If such judgments are too frequent – if without explanation an individual rarely or never acts on practical judgments adducing the presence of a certain type of reason – then we shall conclude that he is merely paying lip-service to the view that it is a reason, and does not really accept it as such (66).

Nagel's reasoning here bears certain similarities to Kant's answer. It is because we are beset by an imperfectly human nature, misled by inclination and clouded by self-interest, Kant tells us, that we do not always act as reason prescribes; and if these "motivational interferences" were to be removed, we would act naturally and effortlessly in accordance with it, as do perfectly rational beings (Ac. 412-414).

But for Nagel to reason similarly is to come dangerously close to begging the question he originally raised for discussion. Nagel answers the question of how motivational content can become motivationally effective by stating that either it can, other things equal, or else the agent does not really accept the justification for the action in the first place. But first, it is hard to see how this could happen on Nagel's account, since this acceptance is built into the judgment. This means that if the agent makes the judgment, she accepts the justification by definition. Second, therefore, this is to claim motivational potential for the justification of an action, other things equal, by definition. It is to claim that, in the absence of irrationality or extenuating circumstances, it is a conceptual truth that an agent can be motivated to act by her acceptance that she has a reason to act. But in order to refute externalism, the claim that the acceptance of a justification for action can motivate action must be defended, not presupposed.

Nagel's reasoning here is analogous to the belief-desire model theorist's claim, that it is true by definition that an agent performs the act he most desires to perform, other things equal. This was the criticism on the basis of which Nagel concluded that reference to such motivated desires could be factored out of a causal explanation of action. If his criticism is valid of the belief-desire model, then it applies equally to the alternative he presents: If the acceptance of reasons for action can motivate action by definition, other things equal, then this principle is merely a conceptual truth and can be factored out of a causal explanation of action. In order to identify genuine causal antecedents of such action, we then must advert to the antecedents of this acceptance.

This objection is not devastating for Nagel's view. Not only acceptings, but believings, cognizings, considerings, reflectings, deliberatings, and concludings are, as we have seen, equally subject to interpretation as

occurrent mental events, and therefore could occupy the requisite causal role, whether they were expressed in well-formed present-tense judgments or not.[11] In these latter cases, it would remain an open question whether, on any particular occasion of action, such an event could precipitate it, or whether other factors might override it, or whether the action, if performed, were overdetermined or not. But these possibilities would be matters for empirical investigation. They would not be motivators by definition.

Nagel then goes on to suggest that if the agent does, however, act on the belief that a present-tense judgment provides a reason for action, then while this belief "does not necessarily imply a desire or a willingness to undertake that action; [while] it is not a sufficient condition of the act or desire", it is sufficient, "in the absence of contrary influences, to *explain* the appropriate action, or the desire or willingness to perform it" (67). Certainly the belief that an action is justified does not *imply* a motivated desire to perform the action. Nor could it – any more than any other mental state by itself, independently of the conjunction of concomitant necessary conditions for action – could provide a causally sufficient condition of the action. Nevertheless, Nagel can claim that under these circumstances, "in the absence of contrary influences," the belief is, *as a matter of fact,* motivationally effective as a precipitating cause of the action. Of course this claim is subject to empirical confirmation. But this is a benefit rather than a disadvantage, for it enables him to avoid the charge of conceptual triviality he leveled against the belief-desire model. If this is the preferred strategy, then Nagel has not finally defended the extraordinary interpretation after all, but rather the ordinary one that accepts premise 2.2.(1) but denies premises 2.2.(2) and 2.2.(3). For on this account it is the occurrent, present-tense *judgment or belief that* one now has a reason for action, i.e. that the action is justified, that is motivationally effective, not those timeless, temporally neutral reasons by themselves.

But Nagel then wants to argue that if a present-tense judgment about what I have reason to do has motivational content, so does the tenseless judgment that it implies. So, for example, if I accept that my present action of arranging alternative transportation to Chicago on May 15 is justified by the airline strike that will occur on that day, then I must accept that Piper's action on May 1 of arranging alternative transportation to Chicago on May 15 is equally justified. This is because acceptance of the present-tense, occurrent judgment implies acceptance of the tenseless judgment plus a [n occurrent] belief that the time of action – i.e. May 1 – is now (68). In both cases, motivational content is provided by my acceptance or acknowledgment that the act is justified (69). The tenseless judgment has motivational content in that my acceptance of it explains why in retrospect, I will want to have acted at the appropriate moment in the past; why at the moment, I want to be acting

[11]I take up this controversial issue in greater detail in Volume II, Chapters II and V.

in the present: and why, in forecasting my future, I will want to act at the appropriate moment later (70). That is, the tenseless judgement implied by the tensed one provides a timeless rather than a dated reason for action.

Nagel's defense of this claim is that acceptance of a dated reason for action without accepting its timeless counterpart entails dissociation from future stages of oneself, which leads to the self-defeating paradoxes of future planning earlier described. This is to say that making only the tensed judgments that embed dated reasons without acknowledging the tenseless judgments – and embedded timeless reasons – they imply would be to exhibit a pathological lack of interest in one's future well-being. But Nagel does not argue, nor should he, that dissociation in turn entails motivational impotence if that dated reason is true for the present moment. What he does point out is that dissociation precludes a dated reason from having motivational content at any other time besides that which it specifies, whereas a timeless reason has motivational content at all of them.

The problems for Nagel's thesis are those already noted. First, if a timeless reason has motivational content at all times, then it cannot explain why I perform the relevant action now without invoking the dated one. The dated reason surely can, but even if it does, as Nagel claims, imply the timeless reason, the timeless reason must somehow reciprocally imply it. For without it, the timeless reason cannot explain the actual action I perform.

Second, in either case, it is unclear which is doing the causal and motivational work in the judgments that embed these reasons: the *acceptance* that the action is justified, or the acceptance *that the action is justified.* In the former case, one could be motivated to act on this basis even if the action were not justified in fact, even if the agent had made glaring mistakes in reasoning, and indeed even if he had not reasoned about the action at all. In this case, the occurrent mental event of acceptance would not be inherently connected to the rationality of the justification accepted, and so would fail to escape the charges of contingency, transience and arbitrariness Nagel leveled against the belief-desire model. In the latter case, Nagel would need to furnish an account of how an abstract propositional object can have motivational influence, which he does not explicitly do. But in either case, an occurrent mental event of acceptance or acknowledgment of a justification is a necessary causal prerequisite to action; hence so is an occurrent, tensed judgment that embeds a dated reason. So it remains the ordinary interpretation of Nagel's project – the one which accepts premise 2.2.(1) but denies premises 2.2.(2) and 2.2.(3), rather than the extraordinary one which rejects premise 2.2.(1), for which the textual evidence is most persuasive.

3. Altruism

3.1. Motivational Action at a Distance

Our task now is to determine whether Nagel's analogous argument for the possibility of altruism offers any remaining considerations that might support the extraordinary interpretation, i.e. that argue persuasively in favor of the rejection of premise (2.2.1) of the belief-desire model of motivation; or whether, on the other hand, Nagel's account of how another's interests can motivate us to act on her behalf relies crucially – albeit implicitly – on mental events of the sort already examined. Nagel defines pure altruism as

> the direct influence of one person's interest on the actions of another, simply because in itself the interest of the former provides the latter with a reason to act. If any further internal factor can be said to interact with the external circumstances in such a case, it will be not a desire or an inclination but the structure presented by such a system of reasons. (80)

As it stands, this definition implies either externalism or motivation at a distance as per the implausible scenario and the extraordinary interpretation, since it permits a person's interest to provide me with a reason to act on her behalf without my being aware of it. Externalism would mean that I might have reason to act but not know it and therefore not act; motivation at a distance would mean that the other's interest might motivate me to act without the intervention of the mental event of my recognizing or acknowledging the other's interest as a reason. Nagel is committed on principle to rejecting externalism and advocating motivation at a distance, so we must suppose him to mean the latter.

Nagel reiterates his rejection of premise 2.2.(1) again for emphasis in the discussion immediately following. He represents the Humean view as asserting that "[w]ith regard to altruism, the corresponding intuition is that since it is I who am acting, even when I act in the interests of another, it must be an interest of mine which provides the impulse," and refutes it as follows:

> The same prejudices are in operation here which have been observed to influence discussions of prudence: the conviction that every motivation must conform to the model of an inner force; the view that behind every motivated action lies a desire which provides the active energy for it; the assumption that to provide a justification capable also of *explaining* action, an appropriate motivation, usually a desire, must be among the conditions of the justification (81).

Clearly, then, Nagel wishes not only to reject the assumption that there must be a desire behind every motivated action, but the further assumption that there must be *any* sort of "appropriate motivation," considered as an "inner force," that "provides the active energy for it." In this passage he explicitly disavows the existence of any such causally efficacious mental event of any kind. As in the case of prudence, he replaces this with a motivated desire

whose causal efficacy is conferred by "reasons which the other person's interests provide" (81), and again the same objection applies: In order for these reasons themselves to be causally effective, I must apprehend them, as per the plausible scenario and the ordinary interpretation. And then it is either my recognition of the rationality of these reasons, or the mere fact of my apprehension of their intentional content that is doing the causal work. In either case, Nagel does not escape the necessity of a tensed and dated mental event that "provides the active energy" for the action it is presumed to motivate. The question that remains to be answered is whether his analysis of genuine reasons as objective and impersonal succeeds where earlier arguments have not.

3.2. Objectivity and Impersonality

Nagel's analysis begins with the distinction between objective and subjective reasons or principles. *Objective reasons* are reasons that express objective value, such that the ends and interests they cite give everyone reason to promote them. Objective self-interested reasons include, for example, personal survival, or happiness; objective non-self-interested reasons might include, for example, fighting injustice, an end which everyone has reason to pursue independently of their personal interests. Nagel's primary concern is with objective self-interested reasons, because these justify altruistic action on the behalf of the agent whose interests are at stake. So, for example, my happiness – a function of my self-interest – is an end of mine which you and others have altruistic reason to promote, just as I have altruistic reason to promote yours and others'. Of course this end will not give everyone reason to perform exactly the same action in order to promote it. I might promote your happiness by tutoring you on the LSATs, whereas someone else might promote it by referring you to a physical therapist. The point is that anyone has reason to promote your happiness, because everyone recognizes the rational value of individual happiness as such.

By contrast with an objective reason, a *subjective reason* cites an interest or end that has value for some agents and not others; specifically, for the agent or agents whose end or interest it is. We may think of a subjective reason as expressing a "for me" value. Thus, for example, neatness at all costs may be a value for me but not necessarily one for you. Hence citing it in an explanation of why I spent an hour arranging my books and papers, sharpening pencils, and doing my roommate's laundry while studying for the LSATs may enable you to understand my behavior without necessarily endorsing it. This example shows that subjective reasons need not cite self-interested or self-directed ends, although of course they might. But they always cite ends and interests that are of value specifically to the agent who holds them, and not to others who do not.

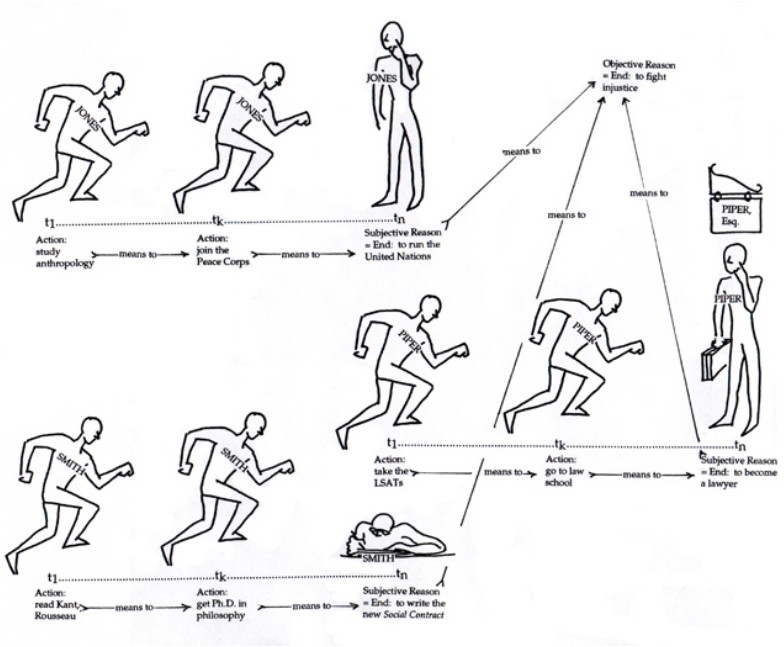


Figure 11. Objective vs. Subjective Reasons

Nagel's thesis is that if an end or interest is a genuine value, then it is an objective value. That is, it provides a reason for everyone to promote it, for one another as well as for oneself. Therefore, the only acceptable reasons are objective. When we cite some consideration as a reason for doing something, we implicitly regard that consideration as having objective value – value not just for us, but for anyone similarly situated. We regard such ends or interests as potentially common rational pursuits for everyone.

This is, in part, a consequence of our impersonal conception of ourselves as one inhabitant of the world among many. Nagel distinguishes between the impersonal and the personal perspective as follows. The *personal* point of view on anything (including ourselves) is the view from the subject's particular vantage point within the world. Expressing the personal point of view usually requires using the first person singular ("I") or token-reflexives such as "this person." Just as in the discussion of prudence a tensed judgment implied a corresponding tenseless judgment plus a statement relating that time to the time of utterance (61), similarly in the discussion of altruism a personal judgment, such as that I need a laptop, implies a corresponding impersonal judgment, such as that Piper needs a laptop, plus a basic personal premise identifying the speaker, such as that I am Piper (103).

Thus the *impersonal* point of view is the view on something or someone irrespective of one's particular relation to it. To view the world impersonally is to view it without regard to one's location in it. Anything that can be said in the first person from the personal point of view can be recast in an impersonal description of that person. Hence the impersonal perspective can accommodate all the facts of the personal viewpoint within it. From this perspective, any statement about a person, including oneself, is such that it can be applied to others under similar circumstances, changing proper names but not propositional meaning. This is the sense in which the impersonal perspective yields the view of oneself as merely one person among others.

To say that we regard our ends as having objective value, i.e. for everyone and not just for ourselves, is to say that our evaluation of a particular end as valuable is an evaluation we implicitly assume anyone could make. Thus our relation to that end is impersonal, not in the dictionary sense of our bearing no personal relation to it at all; but rather in the sense of bearing a relation that is not specific to our personal situation or limited to our own evaluation. For purposes of evaluating the end in question, we are, as individuals, interchangeable with any other equally real inhabitant of the human social world. To view ourselves impersonally, then, is implicitly to view ourselves as subject to universal principles consistently applied to any individual whose situation is picked out by the perfectly general terms of the principle.

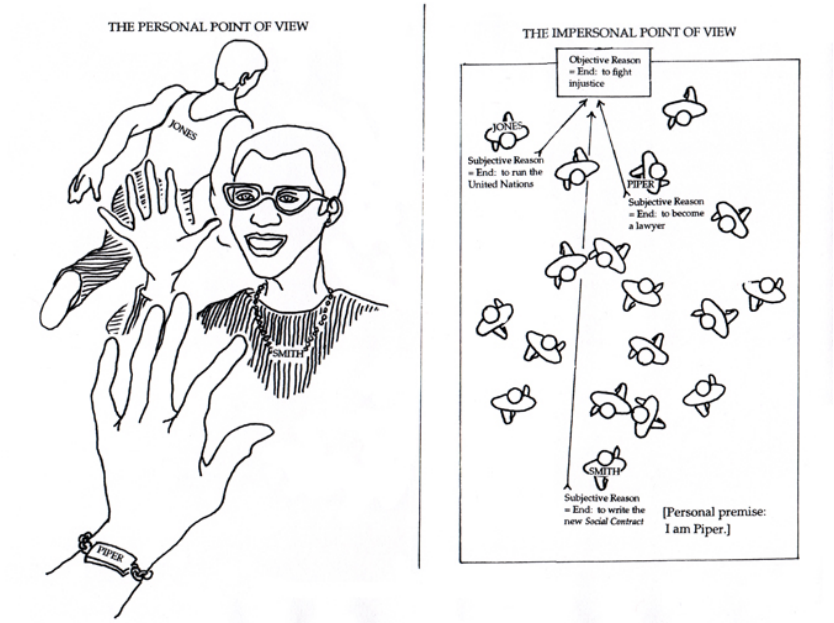


Figure 12. The Personal vs. the Impersonal Points of View

This connection between objectivity, impersonality and universality is the one Nagel needs in order to answer the question raised in Section 1.2, as to whether he could successfully demonstrate the rational inescapability (not the psychological inescapability) of this impersonal conception of oneself as one among many equally real persons. If we think of transpersonal rationality in the weakest possible sense, as requiring consistency and impartiality in the formulation and application of principles or statements of whatever kind, then an impersonal self-conception as one among many possible subjects of universal principles or statements is a self-conception specifically as someone who is a subject of transpersonally rational requirements. And these requirements can be motivationally effective in so far as one truly believes oneself to be the person this self-conception describes. This, I think, is why Nagel emphasizes the efficacy of such principles as a litmus test of whether the agent really "accepts" them: If one experiences no divergence between the person one is and the person for whom conformity to these principles is required, they will be motivationally effective, other things equal. Hence this self-conception is rationally inescapable, not in the sense that we cannot avoid attending to it, but rather in the sense that it expresses a conception of transpersonal rationality as binding on one's thought and action. This is much

closer to the view I defend in Volume II than it is to some of the explicit pronouncements Nagel actually makes.

His project constitutes a rejection, if not a refutation of practical solipsism, understood as the view that one is more real than others. Practical solipsism presupposes that a strictly and narrowly first-personal perspective on the world is the only available one, and that from that perspective only one's own existence – as Descartes' thinking being, perhaps – is ascertainable. From this perspective, universal judgments are at best bogus, assuming they have sense at all, since they can be applied only to oneself. By contrast, Nagel contends that in order to make sense of the ways we actually use language and conceive of ourselves as jointly inhabitants of a shared world, the conception of others as equally real must be assumed at the outset:

> The avoidance of solipsism requires that the *conception* of other persons like oneself (not necessarily the belief that there are any) be included in the idea of one's own experiences from the beginning. This is achieved by a conception which permits every feature of one's own situation and experience to be described and regarded, without loss of content, from the impersonal standpoint (106).

To embrace solipsism is to forego the possibility that universal principles are meaningful, since they can apply only to oneself. It is also to forego the possible that some values are objective, since there are then no others to whom they can supply reasons to act. It is to imprison the interpretation of one's experience within the narrow conceptual constraints of concrete subjectivity; and to confine the governing principles of one's actions to particularistic recommendations as to what this person should do to achieve ends that have value only for her. Nagel rightly argues that we avoid this by adopting the impersonal standpoint on oneself.

3.3. Objectivity and Motivational Content

Nagel contends that from the impersonal standpoint, only objective reasons, and not subjective ones, can motivate action (116-117). To see why, consider the difference between impersonal judgments about subjective reasons and impersonal judgments about objective reasons. Examples of impersonal judgments about subjective reasons would include the following:

> (1) Piper has reason to promote her interests.
> (2) Each person has reason to promote his or her own interests.

3.3.(1), according to Nagel, justifies for Piper her promotion of her own interests, but does not justify for anyone else their promotion of Piper's interests. Since I am merely someone from the impersonal standpoint, 3.3.(1) does not justify for me my promotion of Piper's interests, unless the personal premise, that I am Piper, is added. Similarly, 3.3.(2) justifies for each person

his or her promotion of his or her own interests, but does not justify for just someone that very someone's promotion of each or any person's own interests. Since I am merely someone from the impersonal standpoint, 3.3.(2) does not justify for me my promoting each or any person's own interests, even though I am in fact one of those persons.

By contrast, impersonal judgments about objective reasons function differently:

(3) Everyone has reason to promote Piper's interests.
(4) Everyone has reason to promote a person's own interests.

To register the term, "everyone," is to recognize my own inclusion in its scope of application. Since everyone has reason to promote these interests, I as one among everyone also have reason to do so. Objective reasons can motivate us from the impersonal standpoint because they are reasons for anyone to act on them.

This is Nagel's argument. Its success, however, depends on implicitly restricting the motivational content of an impersonal judgment about a subjective reason to the agent whose subjective reason it is; and it is not necessary to do this. The judgment that Piper has reason to promote her interests, when made to you, may give you derivative reason to promote Piper's promotion of her interests, and perhaps even those interests themselves directly, just in case you think that Piper's having a reason to promote her interests legitimates Piper's promotion of her interests. That she has a *bona fide* subjective *reason* to promote her interests, and not just an unmotivated, valueless desire to do so, may legitimate her promotion of her interests to you, even if you do not share her reason – i.e. even if it is not objective, and therefore do not regard her promotion of her interests as strictly, i.e. objectively justified. So, for example, suppose Piper wants to go to law school in order to become a lawyer – in order, in turn, to fight injustice. Even if you think lawyers are crooks who are incapable even of recognizing injustice, and therefore that Piper's desire to become a lawyer does not justify her spending all that time and money on law school, you may have derivative reason to promote Piper's going to law school if you acknowledge her desire to become a lawyer as a genuine reason, in her eyes, to do what this end requires.

Analogously, Piper might have reason to believe P of herself and thereby give you derivative reason to believe P of Piper, even though you do not share her reason for believing P of herself and therefore do not regard her belief that P is true of her as strictly justified. So, for example, suppose Piper believes of herself that she is impatient, because she gets annoyed when put on "hold" for more than three minutes. You may agree that three minutes is far too long to be put on "hold", and so dispute Piper's characterization of herself as

impatient as too harsh. Yet you may accept Piper's reason for believing she is impatient as evidence for your also believing that she is, in fact, impatient, even though you don't think this evidence by itself is sufficient for concluding that she is (in addition, you notice that she leaves if you are only five minutes late for a lunch date, expects you to return her calls within twenty-four hours, completes your sentences for you so as to get you to the point more quickly, and so on).

In both of these cases, we might say that our judgment that the agent has a subjective reason for action (or belief) gives us reason to *believe in* her – to respect or have faith in her judgment of what is rational for her, and thereby derivatively to support the fulfillment of her projects. What this shows is that there is a difference between having a conscious, motivationally effective reason to act (or believe something) and that act's (or belief's) being strictly, i.e. objectively justified. This corresponds, roughly, to Nagel's distinction between a subjective and an objective reason, such that the former can be distinguished, in turn, from a mere desire. Suppose that a subjective reason is not merely an arbitrary desire but rather has some claim to rational intelligibility. Then an impersonal judgment about subjective reasons can provide one with *some* motivation to act on the agent's behalf, provided that that agent's subjective reason to act itself gives one derivative reason to promote that agent's action or end – irrespective of whether or not one regards that agent's subjective reason itself as strictly, objectively justifying her action. And in the limiting case, in which one is in fact the agent whose subjective reasons are being impersonally considered, one will regard that agent's subjective reasons as at least strongly legitimating her action or end – and so will have considerable motivation to promote them.

Nagel does not intend his account of objective reasons to be a criterion for identifying objective reasons, since he will later go on to say that it

> only places a *formal* condition on reason of whatever kind – the condition of objectivity. No restrictions have yet been placed on the content of those reason and principles which may satisfy this formal condition (125). ... It is clear, in fact, that any catalogue of values can be put into objective form (126).

His reasoning is that any reason for action can be formulated in such a way as to be objective, universally binding, and acceptable from the impersonal standpoint; and that "[w]hatever one may regard as a legitimate goal of action might in principle be regarded as an objective goal – one which anyone had reason to promote" (126).

Nagel should not have conceded this. It implies that for example, Howard Hughes' goal of remaining in bed, unwashed, and narcotized by morphine while screening old movies can be formulated as an objective value that everyone has reason to promote. This is to abdicate Nagel's stated intention of providing a rational criterion for evaluating the moral status of

desires. Nagel should have distinguished between veridical and nonveridical objective formulations of subjective reasons, such that not all objective formulations of subjective reasons are, in fact, of a kind that everyone has reason to promote. So, for example, we may be able to *say*,

> ^Everyone has reason to promote Howard Hughes' morphine addiction.^

But that does not make it true. Since not everyone values Howard Hughes' morphine addiction, not everyone has reason to promote it. Therefore it is not an objective reason, even though it has been formulated as one.

Moreover, Nagel's account of what an objective reason is suggests that the way any particular individual may *regard* or *formulate* her ends is in fact irrelevant to the factual question of whether that end gives everyone reason to promote it. We have already seen that, in order to avoid externalism and the unpalatable implications of the extraordinary interpretation, the notion of a end's "giving everyone reason to promote it" must not mean merely that everyone has a reason even if they don't know it; nor that they have a reason if they know it but are not motivated to act on it; nor that they have a reason if they are motivated to act on it without knowing what the reason is. Instead, an end gives someone reason to promote it if and only if

> (5) it justifies their actions;
> (6) it motivates their actions; and
> (7) it motivates their actions *because they know it justifies their actions*.

That is, it is the recognized rational content of the end that is doing the motivational work. I take 3.3.(5) – (7) to be the necessary and sufficient criteria of rational evaluation of desire implicit in Nagel's account of objective reasons. According to these criteria, there will not be very many objective reasons, but this is as it should be. The promotion of happiness or self-interest, fairness, honesty, the honoring of contracts, mutual aid, fighting injustice, and refraining from harm are some of the familiar candidates likely to appear on a substantive list of goals or values that everyone, in their rational moments, will have reason to promote.

A fuller account of the substantive normative moral theory implicit in Nagel's account would need to say more than Nagel does about the implicitly interpersonal and social enterprise of justification; about what is involved in recognizing rational content as rational; and about the connection between these two and viewing oneself as merely one person among many others. But these issues can be elaborated. There is no need to concede the normative ground to a Humean or Hobbesian "procedure of objectification," as Nagel seems to do in his concluding chapter. Nagel's account contains ample

resources to preserve his distinctions between objective reasons, subjective reasons, and desires that are not reasons at all. And he needs these distinctions in order to explain why practical solipsism is a real condition to be avoided, and not just a linguistic practice that can be evaded through reformulations of principle.

So in order fully to assess Nagel's claim that only impersonal judgments about objective reasons have motivational content, we need to re-examine his claim that such judgments include acceptance of a justification for acting. If I accept a justification for acting every time and only when I judge that I have objective reason for action, then it is hard to see how such an objective reason might ever be outweighed by countervailing subjective ones in fact, and therefore how I might ever justify refraining from altruistic action when an objective reason for such action can be given. Since it always can, it would seem that altruism is not only possible, or even rationally inescapable, but rationally required in all cases. This is far too strong.

3.4. Accepting a Justification

Earlier Nagel argued that first-person present-tense practical judgments that one has reason to act possessed motivational content, understood as "the acceptance of a justification for doing or wanting something" (109), such that that this content is sufficient to explain the corresponding action or desire when it occurs; and that

> it must be present in first-person practical judgments, made from the standpoint of temporal neutrality, and hence also in judgments employing tenses other than the present. I shall now attempt to show by a similar argument that it must be present in impersonal practical judgments as well, and hence in judgments about what others should do (109).

Thus the trajectory of Nagel's argument has been to begin with present-tense judgments made from the personal perspective, then proceed to temporally neutral judgments made from the personal perspective, and finally now to temporally neutral judgments made from the impersonal perspective. His aim has been to show that if the first-mentioned type of judgment can be motivationally effective, then the second can, too, and so, too, the third. From this it follows, according to Nagel, that the third-personal judgments about what others should do that characterize altruistic deliberation can be motivationally effective as well.

He says,

> [W]hat my argument is intended to settle, is whether any motivational content attaches to the impersonal judgment that T. N. has a reason to remove his foot [from under the heel of the man who is about to step on T. N.'s gouty toes], or whether it enters only with the addition of the basic personal premise, 'I am T. N.' (112).

Nagel rightly argues that if the impersonal judgment is supposed not to be motivationally effective, whereas it becomes so with the addition of the personal premise, then since the reason or justification for the action is contained in the impersonal judgment, it cannot be that impersonal reason or justification that is doing the motivational work. This would be what he calls practical solipsism, "[f]or it means that an essential aspect of the first-person judgment, namely the acceptance of a justification, is not present in the impersonal correlate of the same judgment" (113).

There are at least two ways in which this last inference can be interpreted. On one interpretation of it, Nagel is considering the case in which the impersonal judgment that one has reason to act does not, after all, despite his earlier claim (65) include its own acceptance as a justification. This seems right, for we have just seen that once an agent agrees that she has a reason for action, she must then be able to evaluate the force of that reason based on her assessment of its merits. Not even the most rational judgment can be thought to be rationally inescapable in the sense of *compelling* one's acceptance of it as a justification by its content. Since the reason is a propositional object whereas its acceptance as a justification is a psychological event, it is not even clear what it would mean to attempt this.

A stronger interpretation, however, would have Nagel observing that to say that the impersonal judgment has no motivational content, or is not motivationally effective, is, in effect, to say that it has no justificatory force at all; that the reason that is its content is not *capable* of inspiring an agent to accept it as a justification and hence to act on it. And this immediately invites the question, which Nagel himself asks rhetorically, of where this justificatory force could then possibly come from. So we can provisionally agree with Nagel's inference, if this second interpretation is the correct one, that unless the impersonal correlate of the first-personal judgment is supposed to have the same justificatory force, the result must be an irrational and solipsistic dissociation from the rational content of judgments that locate one in the world as one person among many.

But does dissociation from the impersonal standpoint strictly imply solipsism itself? Why not just the sort of externalism expressed in the two following assertions?

> (1) Someone really ought to clean up the garbage in this neighborhood.
> (2) That someone is me.

I might fervently believe 3.4.(1) without its motivating me to act, unless it is accompanied by a belief in 3.4.(2), such that my belief in 3.4.(1) alone contains no motivational content for me. In this case I could make such assertions from the impersonal point of view without their having such motivational content,

because I fail to view myself as merely "someone." I may not grasp that 3.4.(1) implicitly refers to me because I am someone; it may be that I regard me as *me*, and others as "someone." Yet I may have strong feelings of empathy and sympathy toward them, and have a strong sense of their reality. On the one hand, this would be insufficient for impersonal altruism, because these beliefs would lack universality and necessity; but on the other, it would be sufficient to defeat the assumption of solipsism, because it would repudiate the view of others as less real than myself.

So it appears that even if my self-conception is of myself as one person among many others, this by itself does not give 3.4.(1) motivational content. Dissociation from the impersonal standpoint does not necessarily mean I recognize no connection to others at all. It may just mean that my self-conception is grounded in emotional responses to others, rather than in my human and spatiotemporal relation to them. However, Nagel strictly needs only the weaker claim, that if one *does* accept an impersonal standpoint on oneself, this undermines at least the epistemic, if not the ontological presuppositions of solipsism. In this he is surely right.

This second interpretation yields another, less happy implication. For although we may agree that impersonal judgments should have the same justificatory force as personal ones, we must also recall that whether or not an agent actually accepts this judgment as a justification must remain an open question which is not decided by the rational content of the judgment, any more than it is by its impersonal formulation. An agent's *acceptance* of a justification is a first-personal, contingent psychological event, even though the reason that constitutes that justification itself may be embedded in an impersonal, rationally inescapable judgment. And so it now seems clear that it will be possible to raise the same objection about this third stage in the argument that has been raised for the first two: In each case, it seems, it is the dated, first-personal mental event of accepting the reason as a justification that is doing the motivational work, no matter how spatiotemporally neutral or impersonal the rational content of that justification must be.

Nagel has not, then, succeeded in rebutting premise 2.2.(1) of the belief-desire model of motivation. He has not shown that we do not require a present mental event to motivate action. What he has at least suggested is that that mental event need not be a desire. It may be an intentional attitude that is somewhat more susceptible to the persuasive effects of rational content.

3.5. Rational Inescapability and the Kantian Dilemma

It remains to be considered whether Nagel's framework contains the resources for actually solving the Kantian dilemma. Is it possible to retain rational, objective and universal reasons for acting, given the contingent and transient motives on which we act? Are beliefs, considerations, recognitions,

and so forth, any better candidates for "rational inescapability" than desires have been shown to be?

Surely they are at least just as good. The Humean dictum that only desires and not beliefs can cause action seems to be based in the uncritical conception of desire as of what Nagel calls an "unmotivated" desire: one that simply assails us with such force that its causal influence is impossible to withstand. But there are beliefs that simply assail us, whose causal influence is impossible to withstand, too. For example, I cannot help but act as though my dispositional belief in the law of gravity is justified, even though it may not be. I am similarly helpless against my occurrent belief that there is a wasp in my hamburger, however much I may desire – in a nontrivial sense – to take it up and eat it. The inspirational belief that Elvis lives functions with a similar force for some. None of these beliefs happen to satisfy the requirement of rational inescapability. But all of them are competitive with desires in motivational influence, and surpass them at least as candidates for rational inescapability. Beliefs are the kinds of mental entities that *can be* rationally inescapable, whereas desires do not even stand a chance.

Nagel's arguments do not directly address the contingency problem of belief-states as transient and idiosyncratic mental events. But this problem is not necessarily insoluble within the terms of his discussion. Truly held occurrent beliefs include accepting their content as true. An act of acceptance is an occurrent mental event, whereas the content accepted, as an abstract object, is not. The question then becomes a very large one: What causes an agent to accept certain content as true? Nagel would need only a partial answer to this question, namely that *when and only when our rational faculties are properly functioning, the rationality of certain belief-contents itself compels our acceptance of them*, i.e. that the recognition of their rationality is itself motivationally effective (this is recognizable as Kant's answer, to be explored further in Volume II, Chapter V.5 and elsewhere). The question would be not, as Nagel frames it, whether we truly can be said to *accept* a practical judgment that we have reason to act in the event that we are not motivated to act on it (66). It would be, rather, whether we are *thinking clearly* in the event that we are not motivated to act on it. At least some cases of moral failure then would be a demonstration of mental dullness, not one of bad faith. This is an advantage if you believe that at least some people behave badly because they are stupid rather than evil.[12] Of course this thesis would need to be supported by careful distinctions between, for instance, recognizing a belief as rational and mistakenly believing it to be so; between believing P, accepting P as true,

[12]This does not resolve the question of moral accountability one way or the other, since one can cultivate dullness through deliberate habituation, as I have suggested in Chapter VI.3.2. I revisit this matter in greater detail in Volume II.

and recognizing P as true; between recognizing P as true and recognizing P as rational; and so on. But these are not obviously insurmountable tasks.

This is to assign causal efficacy to certain abstract objects, but only under certain conditions: Our rational faculties must function unobstructedly, our attention must be upon the relevant matters, we must be receptive to the assaultive experience of insight, and so on. When these conditions are satisfied, there is a powerful argument to be made that belief-states with rational moral content – for example, that a certain set of considerations constitute a conclusive justification for action – are motivationally effective in causing moral action whenever they occur because of the rationality of their content, i.e. that *rationality itself has a certain pull on us.* Nagel could then stipulate rational content as a necessary but not sufficient condition for the motivational efficacy of the intentional attitude of which it is the object, such that that attitude was motivationally effective only if the content of its object were rational, and not otherwise (in the manner of criteria 3.3.(5) – (7)). In this case, this content would fail to move us to action only if we failed to recognize its rationality, but would invariably move us to action whenever we succeeded. Thus rational content could be a precipitating or contributing but never sufficient cause of action. Also required would be my attention to that content, my recognition of its rationality, and a disposition to act on recognizably rational beliefs.

If Nagel could show this, he would furnish strong evidence for the motivational efficacy of rationally inescapable requirements of altruism on action indirectly. For even if the occurrence of those particular belief-states were themselves sporadic and contingent rather than rationally inescapable, their effect on action once they occurred might be regular, reliable, and perhaps even necessary. Analogously, in the case of theoretical reason, it might argued, it can be true both that our reasoning does not always take the form of *modus ponens* (when we reason at all), and also that when it does take that form and we are reasoning clearly, we always infer Q from P → Q and P, other things equal. Physically acting on the basis of a rational belief that a certain set of considerations constituted a conclusive justification for action would be like the mental act of inferring Q on the basis of the belief that P → Q and P. Thus the rational content of such belief-states would function analogously to the rational content of *modus ponens*. In both cases, it would be the rationality of the content, not merely the occurrent mental event of belief, that was motivationally effective; and in both cases this content could be described as universal without being ubiquitous, and necessary without being compulsive. This is not a possibility that Nagel pursues. But his discussion provides some of the resources on which I draw in order to address it in Volume II, Chapter V of this project.

Chapter VIII. The Problem of Rational Final Ends

I said in Chapter I that a conception of the self consisted in two models. The first was a motivational model. I examined the Humean motivational model itself in Chapter II, the problem of moral motivation it engendered in Chapter VI, and Thomas Nagel's attempt to solve that problem in Chapter VII. I proposed as the second element in a conception of the self a *structural model* that describes the conditions of rational coherence and equilibrium within the self, a model that calls for a theory of rationality to satisfy its requirements. I examined the Humean structural model – the utility-maximizing model of rationality – in Chapters III and IV. In the present chapter I examine the problem of rational final ends that the structural model of utility-maximization engenders for the Humean conception of the self; and I evaluate the attempts of four leading twentieth century Humeans – Frankfurt, Watson, Williams and Slote – to solve it.

Section 1 defines the problem of rational final ends that arises from the Humean model of rationality: that rational criticism or justification of the ends we happen to have – much less revision of those ends in light of alternative conceptions of the good we might develop philosophically – is impossible. Section 2 considers Harry Frankfurt's solution to this problem. Frankfurt distinguishes between free and unfree actions, according to whether or not they are motivated by desires that are themselves the object of higher-order desires that rationally evaluate them. Frankfurt's view implies that an action is free if the desire that motivates it is rational. But it is not easy for a Humean to say what makes a desire rational. Frankfurt's attempt raises the problems simultaneously of self-evaluation and moral paralysis: If we lack desire-independent terminating criteria for evaluating rationally our first-order desires, then on what nonarbitrary grounds do we commit ourselves to any n+1-order desires as themselves authoritative evaluators of our first-order ones? And without such terminating criteria, how can we ever decide what to do with any degree of moral conviction? That we sometimes succeed in doing so suggests that the Humean conception is not adequate to the psychological facts.

Section 3 considers three main responses to Frankfurt's dilemma. Gary Watson proposes a bipartite conception of the self, according to which reason and desire are two independent sources of motivation within the self, as Plato thought. Watson accepts the belief-desire model of motivation. But he also argues that reason is the source of value, of what is genuinely good for the agent, and so enables the agent to assess on which of her desires she should rationally act. But since this Humean/Platonic conception of the self stipulates two independent sources of motivation within the self, it must be an open question not only which *does* take motivational and evaluative precedence on any particular occasion of action, but also which *should*. And in the event that

neither does take motivational and evaluative precedence within the self, the problems of self-evaluation and moral paralysis will not have been solved, but rather exacerbated.

Bernard Williams offers a different bipartite conception of the self to solve Frankfurt's problem, one that also consists in reason and desire as independent sources of motivation within the self. But Williams' variant stipulates exactly the reverse order of priorities from Watson's: that motivational and evaluative precedence is to be accorded those central desires he calls "ground projects," to which considerations of transpersonal rationality are subservient. Williams' demotion of transpersonal rationality is buttressed by Michael Slote's defense of pure time preference, the principle that we should give highest priority to those desires and ends that happen to have the closest temporal proximity to us. Both views assign to such desires and ends the moral importance that substantive criteria for rational final ends would confer. Williams then argues that to sacrifice these desires to the requirements of impersonal moral principle is to alienate oneself from those commitments and attachments that are most deeply expressive and definitive of the self. But detailed scrutiny of Williams' and Slote's claims strongly suggest that they, too, beg both the question of which source *in fact* has motivational and evaluative precedence within the self, and the normative question of which source *should* have it. A personal commitment to rational and impartial moral principles not only does not imply moral alienation; the ability to formulate ground projects, goals, and personal attachments presupposes it. Hence Williams' concept of a centrally definitive ground project does not displace the need for substantive criteria of rationality according to which those ground projects can be reflectively evaluated; and Slote's defense of pure time preference provides no such independent criterion of evaluation. The problem of rational final ends remains unresolved within the constraints of the Humean conception of the self.

1. The Structural Model

In the utility-maximizing model of rationality, desires structure the self in two ways. First, through the distinction into first- and second-order desires,[1]

[1] See Harry Frankfurt, "Freedom of the Will and the Concept of a Person," *The Journal of Philosophy LXVIII*, 1 (January 1971), 5-20. Frankfurt's main thesis is similar to Wright Neely's apparently independent treatment in "Freedom and Desire," *The Philosophical Review LXXXXIII*, 1 (January 1974), 32-54.

Although Neely emphasizes the contrast between the ordinary sense of "desire" as one motive to action among many and the extended philosophical sense that includes all such motives to action, he makes it equally clear that the advantage of the philosophical sense is that it implies means for analyzing all the multifarious motives for the action in terms of "desire" in something like the ordinary sense. Thus he seems to

they determine our evaluation of the other elements of personality: our emotions, beliefs, impulses, and so on. First-order desires are desires for particular states of affairs conceived as external to the self: for unilateral disarmament, for example, or for a slice of carrot cake. Second-order desires are desires for certain first-order desires, hence for their attendant thoughts, feelings and dispositions. Second-order desires are desires that one be (or become) a certain kind of person: they constitute a desired self-conception. For example, suppose I have a central first-order desire for sex, drugs, and rock and roll. This desire may fulfill a second-order desire to be the kind of person who desires such things. Or it may frustrate a second-order desire to be the kind of person who pines only after beauty, truth and goodness. The actual first-order desires that constitute the self either buttress or undermine our desired self-conception. Our second-order desires tell us what that desired self-conception actually is.

According to the utility-maximization model of rationality, the self is structured by its desires in a second way. The importance of instrumental rationality as a defining feature of the self consists in its ability to provide hierarchical order and consistency to the totality of desires one has on any particular occasion: to ensure their mutual consistency with one another, to rank them in order of importance, to schedule a plan for their satisfaction with respect to value, probability, spatial and temporal proximity, duration, and comprehensiveness, and finally to facilitate their satisfaction through maximally efficient action.[2] The structural components of the self are desires, and the rational self is one in which these desires are ordered according to the canons of instrumental reason. Theoretical reason itself is thus a subordinate means for maximizing the satisfaction of our desires.[3]

The structural model of the Humean conception generates a problem about rational final ends because on it, rationality is purely instrumental. It says nothing about which objects of desire are themselves rational, not merely

---

mean "desire" in the technical sense to *cover or substitute for* duties, purposes, intentions, and volitions – as would similarly technical terms like "conation" and "appetition", each of which could be interepreted or analyzed in terms of "desire" in a more ordinary sense – as Neely's examples of duty and wellbeing illustrate. If this interpretation can be carried through, such that each motive to action can be claimed to include a desire in the ordinary sense, then using "desire" in the technical sense to denote all such motives has obvious advantages over terms like "conation" and "pro-attitude."

[2]The Humean conception of the self as structured by the principles of instrumental rationality is explicated in greatest detail in Chapter VII, "Goodness as Rationality," of John Rawls' *Theory of Justice* (Cambridge, Mass.: Harvard University Press, 1971). See especially Sections 63-64 and the bibliography cited there. I discuss this work in Chapter X, following.

[3]David Hume, *A Treatise of Human Nature*, Ed. L. A. Selby-Bigge (London: Oxford University Press, 1968), Book II, 415.

as a means to some further end, but in themselves. It contains no resources for answering that question, nor does it quite acknowledge that the question itself is legitimate. It regards ends such as howling at the moon, counting blades of grass,[4] or lying unwashed in bed in an anonymous hotel room consuming nothing but codeine and old Hollywood movies and being waited upon by lavishly paid employees who also manage one's financial empire,[5] as at most pathological in some psychiatric sense, but otherwise outside the scope of rational evaluation. It thus finds no connection between behavioral pathology and irrationality. The problem, of course, is that there does seem to be a connection. The psychological fact that howling at the moon or counting blades of grass or drinking codeine might be ultimate objects of desire for particular individuals does not excuse them from rational scrutiny, as the Humean rationality model seems to imply. But how can we say what is irrational about these ends, and rational about some others, within the constraints of a rationality model that is silent on what constitutes a rational final end in the first place?

Can we then call on transpersonal rationality to function in a different and noninstrumental capacity? Can it identify any alternative final ends – for example, altruistic or principled moral ones, independent of those objects we in fact happen to desire, that it would be rational for us to adopt? Can it justify the adoption and pursuit of a final end that does not satisfy a desire the agent already has? According to the Humean model of egocentric rationality, it cannot do any of these things. In the end, only the prospect of satisfying an ultimate desire the agent has can justify an extended course of action in the service of the final end which is its object. Relative to this ultimate object of desire, reason can seek out and discover efficient means to that end, and so spark instrumental desires to make use of them. But reason cannot by itself, independently of any such ultimate desire, justify action in the service of principle alone.

So the Humean rationality model implies that, in particular, philosophical reasoning is incapable of articulating persuasively viable alternative conceptions of the good – i.e. conceptions we would be justified in adopting – that diverge from those we already have been conditioned or hard-wired to accept. Philosophical discussion of principles or objects of value – or, for that matter, alternative goals or ends – that do not correspond to those we actually have is pointless, since no such discussion can justify such alternatives. The ends and values we happen to have effectively outcompete any of those proffered as philosophical alternatives, merely because of their

---

[4]This is Rawls' example. *Op. cit.* Note 2, page 432.

[5]Donald L. Bartlett and James B. Steele, *Empire: The Life, Legend and Madness of Howard Hughes* (New York: W. W. Norton and Company, 1979).

seeming self-evidence. Indeed, they may be so deeply ingrained that we unselfconsciously view them as part of the world rather than as part of our value system. Against them, alternative philosophical theories of the good do not stand a chance. The Humean conception of the self implies that normative moral philosophers who devote their energy to elaborating such alternative final ends are best understood as salaried daydreamers. Those who find much to criticize in the desires we now happen to have – or aver, at least, the in-principle importance of rational criticism or justification of those desires – will find this state of affairs less than satisfying.

2. The Infinite Regress: Frankfurt's Humeanism

2.1. Self-Evaluation

The problem of rational final ends is not just about how moral philosophers may be most gainfully employed. It has practical ramifications for the capacity for self-evaluation, as both proponents and opponents of that conception have recognized.[6] The difficulty comes from the assumptions that the self is structured by first- and second-order desires, and that second-order desires provide criteria for evaluation of the motivationally effective desires of the self. The question immediately arises of why we should accept as authoritative criteria these second-order desires. Why should we not subject them, in turn, to the critical scrutiny of third-order desires, and so on, ad infinitum? Frankfurt's answer is that "it is possible ... to terminate such a series of acts without cutting it off arbitrarily," by identifying oneself decisively with one of one's first-order desires. This means that questions regarding higher-order desires are not to arise:

> The decisiveness of the commitment [one] has made means that [one] has decided that no further question about [one's] second-order volition, at any higher order, remains to be asked.[7]

But surely whether any questions remain to be asked about something is not a matter one can simply *decide*. If the state of affairs is unresolved, or suspect, or insufficiently analyzed, then it will raise questions to the discerning observer, regardless of what one has decided; and no amount of mere "decisiveness" will make them go away. If there are no rationally persuasive grounds for halting the ascent to higher-order desires, then the decisive commitment one has made would seem to be arbitrary after all. That I lack the stamina or interest necessary for performing acts of higher-order self-evaluation does not confer authority by fiat on the *n+1*-order desires beyond which I refuse to

---

[6] *Op. cit.* Note 1. Also See Gary Watson, "Free Agency," *The Journal of Philosophy LXXII*, 8 (April 1975), 205-220.

[7] Frankfurt, *ibid.* Note 1, 16.

look, any more than my refusal or inability to consider your point of view settles authoritatively the question of who has prevailed in our disagreement. If an authoritative termination of the infinite regress of orders of desire is to be contrasted with an arbitrary one, we shall need a better reason for doing so than that we are too tired, or unwilling, to press further the hard task of self-evaluation. What we need is some criteria by which to identify certain objects of desire – or ends – as rational in some ultimate and noninstrumental sense. We need criteria of rational final ends.

Hence if the Humean conception of the self is the correct one, we should experience some difficulties in performing the task of self-evaluation. For any set of desires and interests to which I decisively commit myself is likely to seem arbitrary upon reflection. No action can then fully express my self because none can satisfy the desires of my self. And none can satisfy the desires of my self because there are no *n*-order desires with which I can fully identify. The consequence is a desired self-conception attenuated by doubts about the worth and authority of that desire, and so about the action it is assumed to motivate. Leaving unsolved the problem of rational final ends thereby exacerbates the problem of moral motivation already discussed.

2.2. Moral Paralysis

This calls into question the extent to which a self, on the Humean conception, might be motivated to action at all. If the infinite regress of desires prevents one's rational self-identification with any *n*-order set of desires, then there can be no actions to which one can commit oneself wholeheartedly and without reservation – not necessarily because one has conflicting impulses, but rather because the worth of any such impulse is automatically subject to doubt. That I am not in fact left with a continuing case of moral paralysis that vitiates my capacity for decisive and principled action suggests that the Humean model of rational equilibrium does not render accurately the psychological facts.

Some proponents of the Humean conception seem to embrace moral paralysis as a sign of authenticity. Charles Taylor,[8] for example, seems to believe that it is both irresponsible and self-deceptive to presume that one's chosen action might successfully and conclusively quell the stirrings of conscience. He accepts without reservation the implication that dogged and continuing reevaluation of the choices made by the self, and the principled doubt that any such reevaluation is itself adequate, must be permanent features of an authentic self.

[8]in "Responsibility for Self," in A. O. Rorty, Ed., *The Identities of Persons* (Berkeley: The University of California Press, 1976).

Such continuing re-evaluation would require moral paralysis for a subject for whom thought and action were fully integrated, since such a subject would never reach the conclusion that the act in question was justified and so never perform it. A subject with an unintegrated self would perform the act, but without the endorsement of practical reason carried to its final conclusion; without conviction at best, without due consideration at worst.

The idea that there are, and should be, *in theory* no terminating criteria for evaluating the worth of any desire one might have, nor of any action one might undertake, is unsatisfactory. For then either the whole point of ascending to the self-reflective stance of second-order desires in the first place seems to have been lost; or else continual self-reevaluation and principled doubt about the success of one's continuing efforts gain a prominence and normative standing in the life of the mind that is not easy to defend from a moral standpoint. Is it really preferable that we never act out of settled conviction? That we always second-guess our own best moral efforts as to whether they were really good enough? Taylor's conception of moral authenticity seems to court precisely that stereotype of the philosopher as dithering wimp from which the enemies of liberalism obtain so much mileage.

Others may feel no qualms about simply digging in their heels and coupling a forceful assertion of their intrinsic desires – reinforced, perhaps, with whatever resources of force are needed and available to satisfy them – with a bald refusal to give any further justification of those desires. Only the very forceful indeed have this luxury. Nietzsche would approve. But even if such stonewalling or intimidation tactics succeeded in silencing our interrogation of a person's intrinsic desires, they nevertheless would fail to address the question of whether or not such terminating criteria have been met. We are ready to accept such a stance only when they, have, in point of fact, been met: The familiar intrinsic desires for friendship and intellectual stimulation resist further regress, whereas the anomalous or capricious desires to spend one's evenings howling at the moon, or for continuing self-obliteration invite one. The diversity of our responses to such cases may, of course, be purely fortuitous. But it is more likely that the former objects of desire are rationally intelligible whereas the latter are not; and that both sets are susceptible to terminating criteria of rational final ends that the former set satisfies and the latter set violates. However, to explicate these criteria and their relation to the lower-order desires they evaluate requires us to move beyond the scope of the Humean conception of the self. For by definition, the concept of a higher-order desire is insufficient to supply such an explanation; and this is all the Humean model of rationality has to offer.

2.3. Unthinkability

Harry Frankfurt has addressed these criticisms in a more recent discussion, "Rationality and the Unthinkable,"[9] by offering further refinements on the Humean conception of the self. He distinguishes between wanting and willing to perform an action (182, n. 5), and means to make the concept of volition central to his analysis. Indeed, he sometimes uses the term "will" as interchangeable with "intention," (187) which would seem to distance him from the Humean conception entirely. But there are other passages in which he uses the term "will" interchangeably with "want" (184), despite having distinguished them earlier. So his Humean allegiances remain in evidence nonetheless, and I shall treat them as such here.

In this essay Frankfurt approaches the challenge to develop a criterion for identifying those ends of action that are nonrevisable in light of higher-order criteria from the opposite direction. Rather than insist on the "decisive" finality of second-order desires, as he did in "Freedom of the Will and the Concept of a Person," he considers the case of a person – a Utilitarian – for whom no higher-order criterion of rational final ends is unrevisable and for whom, therefore, any end of action must be counted as a real possibility. He describes the resulting condition of moral paralysis this way:

> If the restrictions upon the choices that a person can make are loosened too far, he may become disoriented and uncertain about what and how to choose. ... When he confronts the task of evaluating a large number of addition alternatives, his previously established appreciation of what his interest and priorities are may well become less decisive. ... [S]uppose that now every possible course of action is available and eligible for choice, including those courses of action that would affect the person's preferences themselves. ... But how, then, is he to make any choice at all? What preferences and priorities are to guide him in choosing, when his own preferences and priorities are among the very things he must choose? ... A person like that is so vacant of identifiable tendencies and constraints that he will be unable to deliberate or to make conscientious decisions (177-8).

This is also Rawls' criticism of Utilitarianism, on which Frankfurt explicitly focuses: that in the service of maximizing well-being, a Utilitarian must be prepared to view any end, value, preference or priority as in theory dispensable or revisable – including those which most essentially define her

[9]Harry G. Frankfurt, "Rationality and the Unthinkable," in *The Importance of What We Care About: Philosophical Essays* (New York: Cambridge University Press, 1989), 177-190. Henceforth references to this essay are paginated in the text.

as a distinct and particular individual.[10] In this kind of case, no "decisive" identification with some set of higher-order desires or ends is possible.

To anticipate Bernard Williams' locution (see below, Section 3.2), such an individual lacks personal integrity (178-9). Because any such identification must be conditional on the requirement that the ends in question serve the further end of maximizing utility, no such identification can be unconditional; and therefore no commitment to such ends can centrally define her. Such a person can be said to be alienated from her most centrally definitive ends and values, in the sense that her commitment to them is mediated and vitiated by the necessity of subjecting them to the further requirement of utility-maximization. It is this requirement that provides the terminating criterion of self-evaluation for the utilitarian. But this is no criterion at all. Any substantive end whatsoever – including substantive final ends such as knowledge, happiness, or friendship – can be evaluated by the further criterion of whether or not it maximizes utility in particular circumstances to achieve it. And we have already seen in Chapters III and IV that the concept of utility is either vacuous or inconsistent regardless of how it is interpreted. So Utilitarianism in effect generates the same infinite regress as does Frankfurt's original conception of second-order desires.

This means that there are no ends and values that stabilize such a self in a state of rational equilibrium. The subject fluidly adapts his desires and value commitments to the requirements of maximizing utility under the circumstances in which he finds himself, and changes them as those circumstances do. In the limiting case, such a subject may lack even the minimal psychological consistency that we saw in Chapter IV.3.1 was essential for preserving a sense of moment-to-moment personal continuity. Without independent terminating criteria of rational final ends that determine at what it is rational for that agent to aim, any and all actions and ends can be justified as potential candidates for the maximization of utility.

Now Frankfurt thinks Rawls – and, by extension, Williams, and I – are wrong to draw such a conclusion. He thinks that if a Utilitarian with a specific set of personal values realistically estimates the likelihood of having to revise or adjust certain central desires and capacities as being extremely small,

> [t]here is no reason why [he] should not make to them a commitment that is just as wholehearted as his expectation that he will never encounter circumstances in which maintaining those values would require him to sacrifice well-being (180).

And how wholehearted is this?

---

[10]See John Rawls, "Social Unity and Primary Goods," in Amartya Sen and Bernard Williams, Eds., *Utilitarianism and Beyond* (Cambridge: Cambridge University Press, 1982), 180.

> If his expectation concerning this is unqualified, his commitment also need not be equivocal or at all reserved. ... The possibility that one's expectations are wrong means only that there is a risk in basing a wholehearted commitment on them. It does not imply that taking the risk is either impossible or unjustified (180).

But first, it is not obvious how a Utilitarian could ever come to have such specific values and desires in the first place, unless as mere vestigial remnants of a pre-enlightened moral attitude. How would such values and desires ever get a grip, given the constant reminder that they were to be regarded as conditional on their maximization of utility?

Second, even if they could find a stable place in my overall scheme of values, they would be, in fact, unjustified. If my expectation that I will never have to revise my central desires is unqualified by the acknowledgment that I risk being mistaken in my estimate of the relevant probabilities, then surely I have failed to consider adequately the risk involved in wholeheartedly committing myself to them. And then my wholehearted commitment surely is unjustified, for I have failed adequately to anticipate the likelihood of having to modify it. In Frankfurt's earlier attempt to solve the problem, the vehement emphasis on the decisiveness of one's commitment to one's higher-order desires was insufficient to carry the weight of the argument he gave it. In this more recent attempt, his emphasis on the wholeheartedness of one's commitment suffers the same defect.

Frankfurt offers the concept of unthinkability as a way of understanding what wholeheartedness (or decisiveness) of commitment involves. His argument is that if an agent finds that she cannot bring herself to perform some action that she is in a good position to perform, has reason to perform, and has a desire to perform, then that action violates a wholehearted commitment to some end or value (181). He analyzes this as a case in which the agent cannot will to perform the action in question; and later explains this inability as an *unwillingness* to will performance of the action, and, in the same paragraph, as the agent's not really *wanting* to perform it (184). The inability to "go through with" the action thus stems from value commitments that are so deep and centrally formative of the self that their violation is effectively outside the agent's physical capacity. In this case, the readiness required of the Utilitarian to abandon any such values or ends contingent on their maximization of utility must remain entirely a theoretical matter, even when she justifiably believes that utility-maximization in this instance requires it. Frankfurt seems to think this kind of case demonstrates that even committed Utilitarians may have personal integrity, for there are circumstances in which they seemingly cannot abdicate their most essential commitments even if they want to.

But such a case does not demonstrate this. It demonstrates that one may falsely believe she is prepared to do anything for the sake of utility-

maximization, but discover in the event that there are for her more important considerations that override this one. That is, it demonstrates that one may believe of oneself that one is a Utilitarian yet, when one's convictions are put to the test, turn out not to be one after all. Unthinkability cases show that some agents are, in fact, constrained in their actions by their non-Utilitarian values and ends. But it does not show that a genuine, *wholehearted* Utilitarian can be.

However, the larger problem for Frankfurt's second attempt to meet the infinite regress criticisms is that he no more offers independent, formal criteria for determining what values should finally and authoritatively override others than he did the first time. He offers no terminating criteria for determining which values and ends are rational objects of a wholehearted commitment such that they therefore can receive an agent's fully justified and fully grounded endorsement. He alludes to such criteria by distancing himself from Hume's hyperbolic claim that "'[t]is not contrary to reason to prefer the destruction of the whole world to the scratching of my finger etc." To this Frankfurt rightly protests that a person who really believes this must be crazy; and that surely to say of someone that he is crazy is to evaluate negatively his rationality. But he does not try to explicate the analysis or criterion of rationality according to which we might perform such a negative evaluation.

Instead Frankfurt uses interchangeably the terms "rational" and "reasonable" (185) – a distinction on which Rawls and others have long insisted;[11] equates the irrational with the abnormal (186); asserts a "convergence" among the concepts of the insane, the unnatural, and the irrational (186, n. 8); and observes that "we do sometimes take what we find unthinkable as defining a criterion of normality" (186; also see 187-188). He also argues that

> [r]ationality belongs distinctively to the essential nature of a human being. If we regard a judgment or a choice as opposed to human nature – that is, if it strikes us as unnatural or inhuman – we are inclined to think of it as therefore involving a defect of reason (186).

Frankfurt is of course correct to observe that many people believe that actions or behavior they regard as abnormal, unnatural, or inhuman are for that reason irrational. But this factually accurate report on the vagaries of human judgment does not constitute an argument that what people view as abnormal, unnatural, or inhuman is in fact irrational. Such epithets have been applied to literate blacks, working women, practicing homosexuals, innovative artists, and observant Jews, to name just a few. Frankfurt would

[11] Rawls, *ibid*., 17. Although Rawls later traces this distinction back to Kant (*Political Liberalism* (New York: Columbia University Press, 1996), 48-49), his use of it seems to owe a great deal to the analysis of W. M. Sibley, "The Rational Versus the Reasonable," *Philosophical Review 60* (October 1953), 554-560.

surely agree that the conclusion to irrationality in at least some of these cases would be false.

Finally, he outlines an argument that a person may act rationally when she is acting against her judgment, if her judgment directs her to do what is practically unthinkable for her. We can well imagine such a case: I judge, after long and careful reflection, that in fact I really need only two articles of any kind of clothing (pants, dresses, T-shirts, socks, suits, blouses, shoes, etc.) – one to wear while the other is being cleaned; that it would serve the cause of helping the needy to donate the rest of my clothes to charity; and so conclude with full conviction that I should give the rest of my clothes to the Salvation Army. Yet I find that I cannot bring myself to do it. My emotions, my desires, and my values simply revolt. Frankfurt is surely right to suggest that rationality may be on the side of my emotions rather than my judgment in this case. But he then goes on to conclude that

> it is precisely in the particular content or specific character of his will – which may salubriously lead him to act against his judgment – that the rationality of a person may in part reside (190).

– without, however, suggesting any criterion by which we may identify such rational content. Surely there remains a question to be answered, in the above hypothetical case, as to whether it is my judgment or my emotions that are in fact closer to what rationality requires. Leaving myself with only two articles of each type of clothing may be a very austere, unsociable, and inconvenient way to live. But that does not demonstrate that it is irrational under circumstances in which so many people have no clothes of their own at all.

Frankfurt thus leaves us with a searching analysis of unthinkability as a force that may practically constrain behavior under certain circumstances, and therefore may constrain judgment, desire, or will. This force *may* thus play the practical role that an authoritative criterion of rational judgment, desire, or will would play in guiding, justifying, and motivating action in the service of certain identifiably rational final ends. But it is not the same as such a criterion, nor does it apply coextensively, nor does it suggest what such a criterion would be. In the end, Frankfurt does not supply the rationality criteria that would justify a decisive identification with one's *n*-order desires – or will, or judgments, or intentions – at any level. He thus leaves the fundamental problems of self-evaluation and moral paralysis unsolved.

If there are rational grounds on which decisive identification with one's *n*-order desires (etc.) can be made, this will ensure the authority of the decision to terminate the regress at some particular point in the series, but only by sacrificing the evaluative authority of second-order desires. For whatever the grounds on which we justify our decisive commitment to some set of *n*-order desires, those grounds cannot themselves be desires of any order. If they were, the regress could be reopened, merely by asking for reasons why we should be impressed with the authority of those *n+1*-order

desires. Here it will not do simply to point out that these are the desires we happen to have, or even that these are the final or intrinsic desires which confer urgency on all those that are instrumental to their satisfaction. For that we have desires does not demonstrate that they are non-arbitrary from the perspective of rational justification (suppose, for example, that my most urgent intrinsic desire just is to spend my evenings howling at the moon). *A fortiori*, it does not demonstrate that they constitute rationally authoritative and nonarbitrary terminating criteria of self-evaluation. Hence any such criteria to which we may appeal successfully must be independent, not only of the desires we actually do have, but also of those we should have. For part of the function of such criteria of rationality will be to furnish conclusive and compelling reasons why we should have precisely those desires rather than some others. The Humean model of rationality, even with Frankfurt's more recent refinements, is incapable in theory of furnishing these criteria.

3. Two Bipartite Conceptions of the Self

3.1. Moral Paralysis: Watson's Platonism

Gary Watson[12] has proposed a conception of the self that addresses this requirement. He suggests that we distinguish reason and appetite as two independent sources of motivation, as Plato did. On Watson's view, reason is the source of evaluative judgments about "those principles and ends which [one] – in a cool and non-self-deceptive moment – articulates as definitive of the good, fulfilling, and defensible life."[13] These constitute rational values which are motivationally effective and from the standpoint of which the worth of our motivationally effective desires can be assessed. Since rational evaluations are of the first order too, the infinite regress – with all the attendant problems already enumerated – does not arise.

Or does it? Watson's picture of rational values suggests that the regress is to be blocked by demonstrating that the ends "definitive of the good, fulfilling, and defensible life" are authoritatively justified, i.e. that it would be absurd or irrelevant to raise any further doubts about the rational value of those criteria. This much seems to follow by definition of "defensible." But this characterization thereby begs the question. For we can agree that the rational defensibility of certain final ends renders them immune to the pressure to push the regress of justification one step further. But merely calling them defensible does not make them defensible. Without knowing what Watson intends by "good," and to whom and under what conditions a life is "defensible," there is no reason why my most favored activity of howling at

[12] *Op. cit.* Note 6.

[13] *Ibid.* page 215.

the moon should not be definitive of the "good, fulfilling, and defensible life" for me. And however ready you may be to accept my chosen way of life, surely you are justified in entertaining further doubts about its rationality. If Watson's rational values are truly rational, then we should be able to give persuasive reasons for holding them, and for according them precedence over the promptings of desire. That is, we should have some reason to believe that we are capable of evaluating ourselves *correctly*. Otherwise, Watson succeeds only in shifting the infinite regress from appetitive desires to "rational" values, rather than terminating it.

Watson not only does not furnish such criteria. In fact, he cannot. For in fashioning a bipartite conception of a self that includes two independent sources of motivation, he leaves open the psychological question of which source is in fact authoritative for any particular self, and begs the philosophical question of which source should be. He is concerned to emphasize that the reason-appetite distinction does not commit us to any necessary or inevitable split between reason and desire, since, for example, we may value certain activities such as eating or sex precisely because of the desires they satisfy.

> But the distinction does commit us to the possibility of such a split. If there are sources of motivation independent of the agent's values, then it is possible that sometimes he is motivated to do things he does not deem worth doing.[14]

However, even this much understates the case. For if there are two, mutually independent *sources* of motivation within the self, then surely it must be an open question with which source the agent identifies on any particular occasion, hence which constitutes her self-conception or (in Watson's terminology) "standpoint."[15] Watson seems to take it for granted that an agent

---

[14] *Ibid*., page 213.

[15] Wright Neely makes this point in anticipation of Watson's analysis (op. cit. Note 1, 42). Watson does not use the term "standpoint" as I do the term "self-conception." He means "the point of view from which one judges the world." (216) He doubts the validity of the Humean conception of the self as scrutinizing and evaluating the worth of its first-order desires. Rather, he believes that

> [agents need not usually] ask themselves which of their desires they want to be effective in action; they ask themselves which course of action is most worth pursuing. The initial practical question is about courses of action and not about themselves.

But this seems part of a general plan to throw out the baby with the bathwater. For in denying that we evaluate our first-order desires from the perspective of second-order ones, he seems to want to deny as well that we act self-reflectively at all. But surely one consideration that favors any action we deem worth performing is that it is consistent with actions performed by the kind of person we aspire to be. The "point of view from

must be identified with the values that come from reason, and dissociate herself from any desires or actions that do not conform to them. But this assumption underestimates the role of action as expressive of the self. When I perform genuine action, there is a state of affairs that I envision as its outcome, intend to bring about, and work to bring about. The "I" in the preceding sentence is not neutral between reason and desire. Whichever source of motivation is causing the action is the one that, for that moment, expresses my self. If desire is motivating the action, and reason disapproves of it, then so much the worse, for the time being at least, for reason. And if the conflict persists over the long term, so much the worse for the unity of the self.

Hence the problem of moral paralysis resurfaces in the form of a dilemma for the Platonic bipartite self: Which part of the self ought to have motivational priority on any particular occasion? And who – or what – ought to settle this question? If I act on my desires at the expense of reason, reason can reproach me with incontinence; or, at worst, Aristotelian self-indulgence. If my rational values take motivational precedence over my desires, the approval of conscience may be insignificant in the face of the frustration, regret, and alienation contingent upon ignoring the acknowledged demands of desire.[16] If I am unlucky enough to be torn by equally strong but conflicting tendencies from reason and desire, I may be as fully paralyzed as Buridan's Ass, and for much the same reason. If not, I will be in any case unable to exercise my agency in determining my behavior, and so will suffer the disquieting experience of being propelled into action by forces external to my will, *regardless of the course of action on which I finally embark.*[17] Under such conditions of perpetual internecine conflict, it is a wonder that we manage to do anything at all.

And so for Watson's Platonic conception of the self, the practical problem of moral paralysis is not resolved but exacerbated. This conception fails to resolve the problem because it contains an unexplicated assumption about which feature of the self has rational authority – and therefore motivational priority. Hence his proposed solution to the problem of self-evaluation suffers accordingly. If reason and desire must vie for control of the self as the original picture seems to suggest, then to appeal to rational values to terminate the

---

which one judges the world" is the point of view of a certain kind of self whose capacities for critical scrutiny are exercised as often on itself as on other objects.

[16]Bernard Williams, "A Critique of Utilitarianism," in J. J. C. Smart and Bernard Williams, *Utilitarianism: For and Against* (New York: Cambridge University Press, 1973); "Persons, Character, and Morality," in Rorty, *op. cit.* Note 8. W. D. Falk makes a similar point in "Morality, Self, and Others," in Judith J. Thomson and Gerald Dworkin, Eds., *Ethics* (New York: Harper and Row, 1958).

[17]Frankfurt makes this point about intentional action without agency (*op. cit.* Note 1, page 16), without seeing its implications for the Humean model of rationality.

proliferation of orders of desire is no less arbitrary than it would be to appeal to any appetitive desire to do so. But in the absence of any further, highest court of appeals within which these conflicting demands can be adjudicated, a rationally and morally imperfect agent who nevertheless acts decisively and well much of the time must remain a theoretical enigma.

Thus if we cannot provide, even in theory, some such terminating rational criteria for self-evaluation, it is unclear why we should bother to evaluate ourselves in the first place. Without an authoritative justification of the ends, values and norms on which we both act and rely for criteria of self-evaluation, there is no non-arbitrary reason why we should commit ourselves to those values rather than to some others. Then it is not easy to explain how or why our actions and character should matter, either to us or to anyone else, at all.

3.2. Moral Alienation: Williams' Anti-Rationalism

Bernard Williams has offered the most sustained defense of the position that not only are no authoritative criteria for rational final ends needed; no such criteria are desirable. His view is that appeal or attention to such criteria corrupt moral integrity and promote moral alienation, for they require one to subordinate one's "ground projects" – i.e. one's most centrally definitive final ends – to the requirements of an impartial, universalistic moral point of view that accords those projects no special priority in deciding what one ought to do. I shall call this *Williams' thesis*.

As we have just seen, Williams' thesis and Rawls' elaboration of it constituted the target of Frankfurt's counterargument, just considered, for unthinkability as a terminating criterion of rationality – at least in some cases – that nonetheless would not imply the subordination of one's central ends and values to any external, alienating principle. We have also seen that Frankfurt's counterargument failed because it depended on a case-by-case treatment that declined to specify in general terms when unthinkability is rationally justified and when it is not. Without an explicit, general criterion of rationality, Frankfurt's analysis of unthinkability collapsed into an elaboration of Williams' thesis, rather than standing as an alternative to it.

Williams' thesis does not deny the justificatory function or motivational efficacy of rational principles of morality or justice. Instead it questions their authority to legislate an agent's ends, and their consequent conduciveness to the agent's moral wellbeing. In effect, Williams' thesis is a sustained defense of the claim that final ends do not require rational justification at all. From it I conclude, however, that the moral point of view at which Williams directs his criticisms is either a straw man, or else a foil for a very real problem that is nevertheless unconnected to the enterprise of Socratic metaethics.

3.2.1. Williams' Thesis

3.2.1.1. Ground Projects

What, exactly, is a "ground project"? Williams develops this notion in a series of papers that focus on particular normative moral theories rather than on their metaethical foundations. Consistent with his Humean Anti-Rationalist convictions, his exposition is unsystematic, as it always is with rationally consistent Humean Anti-Rationalists. My avowedly rationalist reconstruction of Williams' thesis attempts to systematize it.

In "A Critique of Utilitarianism,"[18] Williams reproaches Utilitarianism with treating our moral feelings "just as unpleasant experiences," [CU, 103] when in fact they are more accurately "regarded as indications of what [one] thinks is right and wrong." [CU, 103] These feelings, and the "sense of what we can or cannot 'live with',"[19] "partly determine our moral relation to the world, and so cannot properly be understood as "happenings outside one's moral self." [CU, 104] Moral feelings, then, are much more complex and integral to moral agency than Utilitarianism according to Williams acknowledges.

Williams accuses the Utilitarian of a similarly simplistic concept of desire, as consisting solely in "egoistic inclinations and necessities at one end, and impersonally benevolent happiness-management at the other." [CU, 112] But, he argues, one may desire things for oneself, and for "one's family, one's friends, including basic necessities of life, ... objects of taste," as well as "... pursuits and interests of an intellectual, cultural, or creative character." [CU, 110] These latter may be different because "some people's commitment to these kinds of interests just is more thoroughgoing and serious than their pursuit of various objects of taste, while it is more individual and permeated with character than the desire for the necessities of life." [CU, 111]

Now I argued in Chapter VI that all such desires are a species of self-interested motivation; and Williams' characterization of some of them as different, more thoroughgoing and serious, and more individual and permeated with character does not explicitly contradict this. However, each of these desires are what Williams calls "first-order projects," [CU, 110] as is

---

[18] In J. J. C. Smart and Bernard Williams, *Utilitarianism: For and Against* (New York: Cambridge University Press, 1973). Henceforth page references to this essay will be in the text, preceded by CU. I adopt the same convention for other essays by Williams discussed here, as follows: "Persons, Character and Morality, " in *Moral Luck* (New York: Cambridge University Press, 1981) = PC; "Utilitarianism and Moral Self-Indulgence," also in *Moral Luck* = UM; "Morality and the Emotions," in *Problems of the Self* (New York: Cambridge University Press, 1973) = ME; and *Ethics and the Limits of Philosophy* (Cambridge, Mass.: Harvard University Press, 1985) = EL.

[19] Here we find the inchoate origins of Frankfurt's analysis of unthinkability.

one's support of a cause, or those that "flow from some more general disposition towards human conduct and character, such as a hatred of injustice, or of cruelty, or of killing." [CU, 111] None of these desires need include, or be identical with, the pursuit of happiness or pleasure as such. [CU, 112-113] On the view defended in Chapter VI, all of these other-directed desires entail and envision personal satisfaction, and so are no less egocentric than self-directed ones.

A *project*, then, on Williams' thesis, is an object of desire. Objects of desire identifiable as projects range from the egoistic through the social, cultural, aesthetic, and political to the straightforwardly altruistic or benevolent. These objects may be expressive, to varying degrees, of the agent's character and seriousness of purpose, and of his relations to others. Those objects of desire that flow from his moral feelings, e.g. his hatred of cruelty, partly determine his moral relation to the world. Williams sometimes uses the term "project" interchangeably with the term "commitment" [CU, 112, 113] but also describes "commitments" as a special *kind* of project, i.e. "those with which one is more deeply and extensively involved and identified ... [and] which in some cases [one] takes seriously at the deepest level, as what [one's] life is about." [CU, 116]

Projects or commitments thus differ from Williams' version of the Utilitarian concept of desire in three ways. First, the *object* of a project or commitment may be more intentionally complex than a mere feeling, and more psychologically and socially complex than the object of a universal benevolent desire. Second, the *motivation* for adopting a project or commitment may be more complex than a simple egoistic or benevolent impulse. It may also include feelings that are themselves expressive of our moral convictions, and dispose us in various ways towards our own and others' conduct. Third and most importantly, an agent seriously and extensively *identifies* with those projects or commitments that give meaning to her life.

This third aspect of projects or commitments is elaborated more extensively in "Persons, Character and Morality." There Williams introduces the notion of a *categorical* desire as a desire or project which does not depend on the assumption of the agent's existence [PC, 11] – as does, for example, the desire for a sedate lifestyle. Instead, a categorical desire settles the question of whether she will continue to exist or not, i.e. whether she has anything to live *for*. A categorical desire thus may be relatively humble or taken for granted [PC, 12], as might be, for example, the desire to see one's children reach adulthood. Nevertheless, in promising the agent a source of happiness that prevents her questioning the assumption of her existence, categorical desires "propel" her forward into the future, and thus "not only provide the reason for an interest in what happens within the horizon of one's future, but also constitute the conditions of there being such a future at all." [PC, 11]

Categorical desires are ordinarily formed within and by "the dispositions which constitute a commitment to morality." [PC, 12] Nevertheless, the possibility of a radical conflict with morality exists. A categorical desire that

(1) is closely related to an agent's existence;
(2) gives meaning to and a reason for his life in the sense just described;
(3) thus provides the motive force that propels him into the future; and
(4) may radically conflict with morality

is what Williams calls a *ground project*. [PC, 12-13] Ground projects, he tells us, need not be selfish or self-centered. They may be altruistic, or require self-sacrifice on the part of the agent. [PC, 13] Or they may be more like an involvement with one particular other person. [PC, 16] The main idea of a ground project is that the agent identifies with some complex object of desire outside himself, with which he is thoroughly involved, and which gives his life meaning. [CU, 113] To have such projects is to have what Williams calls a *character*.

Williams' notion of identification and thoroughgoing involvement with certain central ends and values is very similar to Frankfurt's notion of a decisive and wholehearted commitment to such ends and values. Both are intended to convey the idea of ends and values that are so deeply embedded in an agent's psychology that they effectively govern her entire existence, whether they are morally or rationally justified or not. So deeply embedded are they, the implication seems to be, that they render the problems of the infinite regress, self-evaluation, and moral paralysis either practically irrelevant or a symptom of pathology.

3.2.1.2. The Moral Point of View

By contrast, Williams understands the notion of an *impartial, moral point of view* as connected with the Utilitarian's higher-order project of maximizing desirable outcomes [CU, 114], and describes it as *sub specie aeternitatis*. [CU, 118] Williams also characterizes the moral point of view as

(1) different in kind from a self-interested point of view;
(2) impartial;
(3) indifferent "to any particular relations to particular persons." [PC, 2]

Moreover, the moral thought that presumably issues from the moral point of view "requires abstraction from particular circumstances and particular characteristics of the parties, including the agent, except in so far as these can

be treated as universal features of any morally similar situation." [PC, 2][20] Finally, the motivations of a moral agent who acts from the moral point of view "involve a rational application of impartial principle." [PC, 2] In this respect, Williams argues, they are different from nonmoral motivations for treating certain individuals differently from others because of one's interests or feelings towards them. Moral motivation does not thereby exclude such treatment; but it does claim a "special dignity or supremacy" which nonmoral motivation seems to lack. Moral motivation and the moral point of view thus involve a "detachment ... from the level of particular relations to particular persons, and more generally from the level of all motivations and perceptions other than those of an impartial character." [PC, 2] In particular, they therefore involve a detachment from what Williams has defined as the agent's character. [PC, 14]

Clearly, the notion of impartiality plays a central role in Williams' conception of the moral point of view and the motivation that characterizes it. But it is not easy to understand the use he makes of this notion. *Impartial* means "unbiased, unprejudiced, just, fair, or equitable," according to the *OED*; but Williams seems to mean something different. His talk of detachment, indifference to, and abstraction from particular relations, persons, circumstances, characteristics, feelings, motivations, and perceptions suggests that he means by "impartial" what the *OED* defines as *impersonal*, i.e. "having no personal reference or connection." And indeed he seems to dismiss Rawls' suggested distinction between them. [PC, 5] By an "impartial moral point of view," then, Williams would seem to mean a point of view of the sort Thomas Nagel elaborates (see Chapter VI.3.2, above), from which one ignores any personal reference or connection the object viewed may have to oneself as a social and affectional agent. Henceforth I shall refer to this notion as an *impersonal moral point of view*, for the sake of precision.

The principles or theory that govern one's behavior and relation to others from this point of view can be described as *universalistic*, to use Williams' own, more recent terminology, if they are

> (4) *universal* in that they apply to all subjects and objects similarly designated as within their scope (.e.g. persons, moral agents, sentient beings, etc.);

[20]In this respect, the difference between a Kantian and a Utilitarian view must be one of degree rather than, as Williams seems to suggest, one of kind. For presumably the empirical information on the basis of which the Utilitarian calculates the best action in a particular case will yield the same calculation in other, morally similar cases. What seems to distinguish Utilitarianism is the large amount of empirical information relevant to determining the moral similarity of different cases.

(5) *general* in that they include no proper names or definite descriptions;[21] and

(6) *impartial* in the strict sense, in that they accord no special privilege to any particular agent's demands or requirements that cannot be justified under requirements (4) and (5).

Further evidence for this reading can be culled from Williams' later *Ethics and the Limits of Philosophy*. For example, he characterizes a reflective question as *general*, "because it asks how to live," [EL, 19], and *timeless*, in that "it invites me to think about my life from no particular point in it." [EL, 19] A reflective question that can be asked by or of anybody is one that is asked from, and leads us to an answer from, the "impersonal standpoint." [EL, 20] This characterization of the impersonal standpoint squares well with Nagel's conception, as well as with the dictionary definition of impersonality cited above.

Similarly, Williams describes the "outside point of view" of one's ethical dispositions as one in which one may try to "abstract [one]self totally from those dispositions, and to think about [one]self and the world as though [one] did not have them." [EL, 51] This would seem to be a case of regarding oneself and the world from an impersonal standpoint. Later, Williams characterizes "the standpoint of impartiality" as that from which an agent reflects on herself qua agent, hence sees herself "as one agent among others." [EL, 65] The agent thus "stands back from [her] own desires and interests, and sees them from a standpoint that is not that *of* [her] desires and interests. Nor is it the standpoint of anyone else's desires and interests." [EL, 66] Again, this is a standpoint on oneself "having no personal reference or connection."

Finally, Williams proceeds to argue that there is a radical asymmetry between theoretical and practical deliberation that Kant's account of rational freedom ignores. By contrast with practical reasoning, in which "the first person is not derivative or naturally replaced by *anyone*," [EL, 67] theoretical reason is essentially third-personal. For the "I" occurs only incidentally in such factual statements as "Wagner never met Verdi." Here the prefix "I believe ..." is a "derivative, merely reflexive counterpart to the thoughts that do not mention me. I occur in them, so to speak, only in the role of one who has this thought." [EL, 67] From the premise that, unlike theoretical deliberation, practical deliberation is essentially first-personal, Williams infers – contra Nagel – that no "necessary impartiality" can be imposed on it: "[P]ractical deliberation ... involves an *I* that must be more intimately the *I* of my desires than [Kant's] account allows." [EL, 67] So Williams seems to want to claim that the "I" who has these desires is personal, whereas the "I" who has these thoughts is impersonal. The idea would then be that to impose the

[21] Rawls, *op. cit.* Note 2, Section 23.

requirement of impersonality on the viewpoint of an agent engaged in practical deliberation produces a psychologically unnatural distortion of what is essentially a personal endeavor.

This last argument is a peculiar one to level against a philosophical view that centrally includes the thesis that factual propositions would be unintelligible unless they could be ascribed at least implicitly to the "I" who has the thoughts they describe.[22] Kant's thesis would seem to suggest that the "I" of practical reason and the "I" of theoretical reason must stand or fall together with respect to their personal connectedness. But perhaps Williams' silence on this major thesis of Kant's *Critique of Pure Reason* is intended to dismiss rather than overlook it. In any case, Williams' remarks here support the suggested interpretation of his notion of impartiality as, more precisely, *impersonality* in the dictionary sense, since factual statements *sans* intentional operator lack personal reference or connection. He seems to want to rebut Nagel's analysis of practical reasoning, but without giving it the degree of scrutiny Nagel's complex analysis requires.

Williams' claims regarding the universality of moral principle support the connection described above and in Chapter VII.3.2, between the impersonality of one's standpoint and the universality of the principles that then may be supposed to govern one's behavior. For example, he argues that the Kantian ideal of moral theory is too universalistic to fit our ethical intuitions: "We can less and less appropriately rely on those intuitions that belong distinctively to the local *we*, since the theory is now to be a theory for an *us* that includes agents existing far away from our local folkways;" [EL, 103] that "the universalistic perspective will not determine the content of the ethical theory for a given group;" [EL, 104] and that the notion of a "universalistic standpoint" involves a mistaken "Platonic assumption that the reflective agent as theorist can make himself independent of the life and character he is examining." [EL, 110] This would be to examine critically one's dispositions "from the outside, from the point of view of the universe." [EL, 110] But the point of view of the universe, Williams thinks, just is not the point of view of human beings. [EL, 118] These comments conjointly imply that

> (7) the moral point of view is the "outside," "third-personal," *im*personal point of view of the universe, that includes no personal reference or connection to the agent whose view it is;
>
> (8) the principles of conduct that govern this point of view are universal, general, and impartial in the sense explained; and

[22]Immanuel Kant, *Kritik der Reinen Vernunft*, Herausg. Raymund Schmidt (Hamburg: Felix Meiner Verlag, 1976), A116-117, 123-124, B 131-140, 155-159, *passim*.

(9) the moral point of view and its governing principles are therefore detached from the ground projects, character, and human point of view of individual human agents.

3.2.1.3. Integrity and Alienation

What, then, are "moral integrity" and "moral alienation"? The foregoing conclusions enable us to answer this question relatively straightforwardly. *Moral integrity* is a state of the agent in which the agent views the world from what Nagel would call the subjective perspective, i.e. that of his own projects and ground projects. From this perspective the agent is disposed to act on the patterns of feeling and desire that constitute those projects, and does so unselfconsciously, i.e. without formulating universalistic principles to which he intends his behavior to conform. Williams articulates the concept of moral integrity most explicitly in "Utilitarianism and Moral Self-Indulgence." There he characterizes integrity as neither a virtue nor a disposition nor a motive. "It is rather that one who displays integrity acts from those dispositions and motives which are most deeply his, and has also the virtues which enable him to do that." [UM, 49] In this sense, moral integrity may conflict with universalistic moral principles, such as those definitive of Utilitarianism. [UM, 51] And in *Ethics and the Limits of Philosophy*, he describes with approval Aristotle's ideal of the life of practical reason as including "certain excellences of character or virtues, which are internalized dispositions of action, desire, and feeling [EL, 35] ... It is an intelligent disposition. It involves the agent's exercise of judgment, ... and so it is not simply a habit. It also involves favorable and unfavorable reactions to other people, their character and actions." [EL, 36] These last passages are important, for they suggest that Williams' concept of moral integrity does not involve a lack of rational reflection, just an absence of "universalistic principles." [UM, 52] Rational but nonuniversalistic deliberation presumably would deploy particular instances of universal principles of means-end reasoning such as,

In order to get into law school, I will have to pass the LSATs.

The rational reflection of a morally integrated agent is characteristically directed at the object of his desires or projects, not at himself as subject of them. [UM, 48; EL, 10]

Moral alienation, by contrast, is a state of the agent in which the agent views the world and her projects from the impersonal moral point of view, is disposed to sacrifice them to the requirements of moral principle, and does so with a self-conscious awareness of conforming to those requirements. Thus we are alienated from our moral *feelings* if we "come to regard those feelings from a purely utilitarian point of view, that is to say, as happenings outside one's moral self." [CU, 104] We are alienated from our *actions* if, "when the

sums come in from the utility network which the projects of others have in part determined," we suppose that we "should just step aside from [our] own project and decision and acknowledge the decision which Utilitarian calculation requires." [CU, 116] We are alienated from *others* to whom we have deep attachments if our reactions to them are motivated by "one thought too many," [PC, 18] namely that moral principle condones (or disapproves) those reactions. [Also ME, 227] Moral alienation is, for Williams, the consequence of a *purist* moral attitude that insists on "abstracting the moral consciousness from other kinds of emotional reaction or social influence." [EL, 195] For under these conditions, impersonal moral consciousness detaches us both from our personal responses and thereby from ourselves. Now we saw that Nagel reproached the personalism and subjectivism of the Humean Anti-Rationalist with the charge of solipsism. Williams's response is to reproach the impersonalism and objectivism of Kantian and Utilitarian Rationalists with the charge of moral alienation. Thus each levels the charge of dissociation against the other.

3.2.2. Moral Theory

3.2.2.1. Structure

First some preliminary remarks on the structure of Williams' thesis. As we have seen, Williams' concept of a desire as a *project* is reformatory in nature. Desires, he wants to say, are considerably more various and complex entities than many moral theories have recognized. He does not question the basic Humean dictum that all action is motivated by desire; thus, for example, his argument that not every desire aims at pleasure because if it did we could make no distinction between ethical and hedonistic motivations [EL, 49-50] presupposes that desire is the only sort of psychological entity that can motivate at all. But he does claim for desires or projects as "first-order motivations" [UM, 46] that they are not "dissociated" from complex thoughts and judgments we may make about their objects [UM, 52; EL, 36] Thus, for example, he characterizes the Aristotelian account as one in which "my reflection, even if it is about my dispositions, must at the same time be expressive of them. I think about ethical and other goods *from* an ethical point of view that I have already acquired and that is part of what I am ... [i.e.] someone in whom the ethical dispositions he has acquired lie deeper than other wants and preferences." [EL, 51][23]

Opposed to desires of this more complex kind is what I have described as the Nagelian impersonal moral point of view, i.e. the view we take of ourselves and others *sub specie aeternitatis*, as one agent among others,

[23] Also see EL, 112 and 116, and the discussion of "thick" ethical concepts at 140-142, 200.

independent of any personal relations or connections, from which Williams supposes "universalistic" principles to be generated. Universalistic principles are appropriate to scientific theory, because it "looks characteristically for considerations that are very general and have as little distinctive content as possible, because it is trying to systematize and because it wants to represent as many reasons as possible as applications of other reasons." [EL, 116-117] By contrast, the "critical reflection" appropriate to the enterprise of ethics "should seek for as much shared understanding as it can find on any issue, and use any ethical material that, in the context of the reflective discussion, makes some sense and commands some loyalty." [EL, 117] This is because, as we have just seen, such reflection must express our motivational dispositions and projects rather than detach us from them. Hence we can understand Williams' thesis as admonishing us not to assume a stance that is, in fact, part of our selves, but that leads us astray when we are trying to solve practical questions.

Williams' thesis thus depends on a bipartite conception of the self that, like Watson's, draws upon the Platonic distinction between reason and desire. On one side are the impersonal point of view and the universalistic principles expressive of it, appropriately deployed in scientific but not ethical theorizing. On the other side are the more complex and reflective desires that Williams calls "projects", from the perspective of which we appropriately survey the world and engage with it ethically. Thus unlike Plato and Kant, and twentieth century philosophers such as Thomas Nagel, but much like Hume and twentieth century philosophers such as Richard Brandt, John Rawls, and Harry Frankfurt, Williams claims authority for desire over theoretical reason in the structure of the self. Williams' contribution to this Humean conception of the self is to have pointed out that if a psychologically adequate concept of desire cannot depict it merely as an "original ... modification of existence ... [that] contains not any representative quality,"[24] then there is no reason whatsoever to suppose that objects of desire must be limited to personal gratification on the one hand and universal happiness on the other.

Theoretical reason, in turn, according to Williams' thesis, has two parts: the impersonal point of view and the universalistic principles that govern that point of view. The passages already examined show that Williams – like Nagel – thinks of these two as necessarily interconnected, but this is mistaken. An agent may adhere to universalistic principles without adhering to an impersonal point of view, and may adhere to an impersonal point of view without adhering to universalistic principles. If either of these possibilities can determine ethical behavior in a psychologically and socially healthy way, then Williams, like Watson, has effectively begged the psychological question of

[24] David Hume, *A Treatise of Human Nature,* Ed. L. A, Selby-Bigge (Oxford: The Clarendon Press, 1968), 415.

which aspect of the bipartite self is in fact central to the structure of actual selves, as well as the normative philosophical question of which aspect should be.

3.2.2.2. Personal Investment (Attachment Revisited)

Consider the first possibility, that an agent adheres to universalistic principles without adhering to an impersonal point of view. Suppose you are personally invested in a moral theory that contains various universalistic prescriptions of fairness, sympathy, honesty, compassion, and so forth. Also suppose that there are clear psychological reasons for your personal investment, deeply rooted in your personal history. Perhaps, in addition to having had a sound moral upbringing, you discovered early on that these deeply instilled principles were your only resource for coming to terms psychologically with repeated personal injustice, or confusing or threatening personal encounters. Suppose further, then, that your investment in these principles informs your entire social and personal life. You try to do what is right, to be fair and honest in your dealings, to understand and sympathize with others, and to respond to them compassionately and without prejudice, as your moral convictions prescribe. These principles also inform your private life: You attempt to secure and maintain good physical health, to live modestly but tastefully, and not to deceive yourself about who you are or to what you aspire. The moral prescriptions that guide this conduct are *universal* in that they apply not only to you, but to all rational human agents. They are also *general* in that they include no proper names or definite descriptions. Moreover, they are *impartial*, for they require you to accord no special privilege to your personal requirements, merely in virtue of the fact that you are the agent whose behavior you are evaluating. Of course your moral theory includes provisions for different circumstances and social relations, for example that the elderly deserve special respect for their wisdom and experience, that one has special obligations to family and loved ones, and so on. Nevertheless, it applies universally, generally and impartially, for there is nothing in it tailored to fit your particular situation.

Now if this moral theory entailed an impersonal or dissociated point of view, then since this point of view is the symptom of moral alienation, we would be forced to conclude from your overriding investment in this moral theory that you were alienated from those of your central desires and ground projects thus overridden. But surely you are so alienated only if your personal investment in your ground projects outweighs your personal investment in the moral theory with which they may conflict; and surely this is a moot question. I shall say that an agent A is *personally invested* in some state of affairs x if

(1) x's existence is a source of personal pleasure, satisfaction, or security to A;

(2) x's nonexistence elicits feelings of dejection, deprivation, or anxiety from A; and

(3) these feelings are to be explained by A's identification with x.

A identifies with x if A is disposed to identify x *as* personally meaningful or valuable to A. This definition of personal investment is the same as the definition of attachment in Chapter II.2.4. Renaming it "personal investment" here highlights that dimension of attachment that involves giving over some part of oneself to the object of attachment, with the expectation of return or reciprocation from it.

According to these criteria of personal investment, you are alienated from your central desires and ground projects if, because you identify with them, sacrificing the possibility of their realization when your moral view so prescribes elicits the feelings of profound loss described by (2). You are alienated from your moral theory, on the other hand, if, because you identify more completely with this theory, these feelings of loss result from pursuing your central projects at the expense of your moral theory. For in this case, it is your moral theory that is the focus of your personal investment, not your desires and ground projects.

Williams does not speak directly to the importance of something like the criterion of person investment in determining when, whether, or from what a person is morally alienated. He seems on the one hand to recognize it, by referring to that from which he presumes agents to be alienated as "commitments," and that with which they "identify." On the other hand, Williams views as evidence for an agent's moral alienation his regarding his desires and ground projects with detachment – i.e. *without* these feelings of loss. But it does not follow from the fact that the abdication or sacrifice of certain things with which we identify fails to elicit profound feelings of loss from us that we are alienated from those things. For they may have been of only peripheral importance to us to begin with.

Consider another, analogous conflict between reason and desire. Suppose you have a settled, long-standing, and recurring desire to smoke. You also hold wholeheartedly the universal, general and impartial conviction that it is both unhealthy and inconsiderate to others to smoke. Although you are frequently tempted to smoke, the force of your conviction usually enables you to resist this temptation. Your unqualified personal investment in the view that smoking is a moral evil leads you to fear your desire to smoke, and to anticipate its onset with anxiety at the possibility that you may give in to it. When it occurs, you are simultaneously torn by craving for a cigarette, and by self-disgust at your inability to rid yourself of this craving once and for all. When the desire passes, you rejoice, relax, and hope that you have seen the

last of it. Of course you experience temporary feelings of deprivation, dejection, and anxiety at not smoking when the desire to smoke overtakes you. But these feelings are to be explained by your physical addiction to smoking, not your identification with it. Though you often desire to smoke, you have no personal investment in smoking. If you have no personal investment in smoking, it is inappropriate to describe you as alienated from your desire to smoke, even though that desire may be a long-standing, central, and powerful one.

If one is not commonly assumed to be alienated from the desire to smoke when one withstands it in favor of the universalistic prescription not to smoke, it is unclear why one should be assumed to be alienated from any other desires when one sacrifices them in favor of a universalistic prescription to, say, treat others fairly. Williams' supposition of an agent's personal investment in his projects at the expense of rational principle needs to be demonstrated, not taken for granted.

Williams might reply to this that a closer scrutiny of such apparent counterexamples would reveal that one's investment in fact is in the moral content of the theory, i.e. that *one* do what is right, fair, compassionate, etc. and not in the theoretical or universalistic formulation of that content. But this might be difficult to demonstrate. For, first, part of the appeal of aspiring to conform to such universalistic prescriptions may well be their depiction of oneself as a member of a larger rational community, self-governed by principles shared in common. If one is personally invested in this conception, then to require conformity to these prescriptions only of oneself and not others would be to invite the feelings of profound loss described above (thus are high-minded misanthropists born). To invest oneself in this conception of rational community is not necessarily to abdicate the personal perspective for an impersonal one, any more than I do by investing myself in my conception of my family. To do the latter is at least to identify myself as a member of my family; but this is to see myself from the perspective of *my self* as I identify it, and not from some other perspective. Similarly, one would think, if I identify myself as a member of the rational community.

3.2.2.3. Universalistic Principles

In fact, central desires and ground projects can take *prima facie* motivational precedence over rational principle only if rational principle takes *de facto* precedence over desire. In order for us to pursue our desires at the expense of Williams' moral point of view, we must have beliefs about what those desires are. Call these *desire-identification beliefs*. According to the representational analysis of desire offered in Chapter II.2, a desire is a particular conventional and theory-laden object of conscious representation. A desire-identification belief is a species of conscious representation that identifies such an object as having certain specific properties.

In order for desire-identification beliefs to facilitate rather than obstruct the pursuit of our desires, at least some of them must be logically consistent in content. In order for them to reidentify a desire as the same from one occasion to the next, distinguish among conflicting desires, and rank desires relative to one another in order of priority, at least some of these desire-identification beliefs must be general in the sense explained in Section 1.2, above. For if each of them contained either proper names or definite descriptions, we would lack that generic concept of a desire that enables us to identify each of them as being of a certain kind, comparable to or contrastable with others.

But desire-identification beliefs that are both general and logically consistent apply impartially to all the states of affairs to which they refer, without regard to whose state of affairs they designate. Thus, for example, the proposition that a certain kind of agitation in the presence of a cigarette is a cigarette-craving applies impartially to my own as well as others' symptomatic agitations. In order for me to identify my desire for a cigarette when and if it occurs, I must believe this general proposition consistently. Similarly with the desire-identification belief that cigarette cravings are unhealthy. And similarly with the desire-identification belief that the desire to live an addiction-free life is worth pursuing. To be moved by desires about which we had no such transpersonally rational beliefs, or by desires about which the beliefs we had failed the transpersonal rationality criteria of generality, impartiality, and consistency, would be to behave blindly and reactively, without conceptual self-awareness. That we usually find such reactive behavior cause for concern or modification confirms the suspicion that the alternative to impartial rationality is not moral integrity but conceptual oblivion. A considered commitment to one's central desires and ground projects presupposes rather than pre-empts transpersonally rational principle.

One might protest that this notion of rational principle appears to be much broader than the one Williams meant to target. For it is unlikely that Williams would want to deny the compatibility of moral integrity and abstract conceptual thought. Williams' criticism is presumably intended to address certain narrower, standard moral theories – specifically Kantian and Utilitarian ones – that he finds deficient in some respect. But the deficiency cannot lie in the universalistic or impartial character of these theories. For any theory or concept applies impartially to all subjects designated as within its scope, including Williams' own.

However, the impartial application of Williams' own principles renders them self-defeating. This is because Williams' criticism implies either that we should either (a) act out of a commitment to our central desires and ground projects, not from the universalistic prescriptions of theory; or else that we should (b) avoid theorizing about our moral condition, alienated or otherwise. Since (b) would immediately obviate the point of advancing Williams' thesis

in the first place, assume for the moment that (a) is the correct implication of his criticism. Since (a) applies impartially to all subjects addressed by the imperative, it accords no special weight to the desires and projects of any particular subject who holds it. It requires that each of us evaluate our motives and behavior impartially, with an eye to their conformity to (a). Hence it appears that (a) is equally susceptible to Williams' critique of moral theory. That is, it would seem that I am alienated from my central desires and emotions by virtue of my commitment to (a), from the detached perspective of which I survey the variety of my motivational states, in order to determine which should cause action. For it is only as the result of successfully withstanding the scrutiny of the detached perspective expressed by (a) that my desires and projects are legitimated as motives for action. Williams' notion of alienation would seem, then, to be an unavoidable consequence of thinking conceptually, in the universal, general and impartially applied terms which language necessitates, about the particular actions we take.

This in turn implies that (b) can have the desired outcome only if we fail to heed it. We can follow this suggestion only by not following it deliberately.[25] In some moral theories, this apparent impasse is often resolved by appeal to an Aristotelian theory of habituation, according to which we are to cultivate certain dispositions in ourselves by repetition, imitation, and any other available behavior modifiers, so as to reach the point at which we are naturally and reflexively inclined to act in the prescribed way, without having to be motivated consciously by the prescription. Analogously, for example, we might be happier by cultivating the dispositions of generosity, cheerfulness, and curiosity, than by continually cogitating about which action available to us would maximize our happiness.[26]

However, both (a) and (b) present special obstacles to this program. For the attempt deliberately to cultivate the dispositions they prescribe is also self-defeating. In order for me to conform to (a), I must conform to (b). But in order to cultivate the disposition to conform to (a), i.e. to act out of a commitment to my central desires etc., I must be prepared to behave in that way whenever the occasion for practice arises. In order to attain this state of psychological readiness, I must think about it – i.e. violate (b). In order to know that I am succeeding in cultivating the favored disposition, I must also think about it. And every time I think about it, thus violating (b), I am undermining the disposition I seek to cultivate, i.e. (a).

---

[25]This argument is defended in Susan Wolf, "Moral Saints," *The Journal of Philosophy LXXIX*, 8 (August 1982), 419-438.

[26]Henry Sidgwick advocates this strategy in *The Methods of Ethics* (New York: Dover Press, 1968), Book II, Chapter II, Sections 2-3, pp. 136-40.

It appears, then, that in order for me to cultivate this disposition successfully, I must do so unintentionally. In order not to undermine it, my program of moral self-improvement must consist in acting blindly as I please, comforted by a self-serving Utilitarian argument that it is at least possible that my behavior may have the outcome I wish. Adherence to Williams' theory requires one deliberately to reject it. With a target of criticism as inclusive as that of universalistic rational principle, it is difficult to avoid shooting oneself in the foot.

This conclusion suggests that to characterize the rational perspective as alienated because it prescribes general and impartially applied norms of behavior – which of course describe particular states of affairs under particular circumstances – is no more convincing than it would be to characterize one's general desires for love, friendship, and physical comfort as alienated because they do not as such exhaustively specify the particular states of affairs that contingently fulfill them. But Williams' thesis commits him to this. It seems that on his view, we are caught between moral alienation and conceptual oblivion. So it could not be, in fact, the universalistic formulation of the content of moral theory as such that is objectionable. For any principle that lacks a proper name or definite description applies universally, generally, and so impartially in the sense defined to all subjects designated as within its scope, including Williams' own. There is reason to doubt that universalistic principles as such engender moral alienation, and hence that they are necessarily interconnected with the impersonal point of view.

3.2.2.4. Slote on the Rationality of Pure Time Preference

We have seen that Williams' attack on universalistic moral principles had particular relevance to Nagel's earlier defense of altruism. But Nagel also defended prudence, and prudence is governed by universalistic principles just as thoroughly. Fellow Humean Anti-Rationalist Michael Slote targets Nagel's universalistic defense of prudence in the same way that Williams targeted Nagel's case for altruism. But in his argument against the irrationality of pure time preference,[27] Slote confronts the same dilemma of self-defeat as did Williams.

Slote formulates the view he means to target as that which claims that "different (properly articulated) times of life are of (roughly) equal importance in determining the goodness of lives" (13), and identifies Thomas Nagel, John Rawls, Amartya Sen, Charles Fried, and Henry Sidgwick as all proponents of this view. Against them, Slote wants to argue that "we typically and naturally think of some times of life as more important than others" (13).

[27] Michael Slote, *Goods and Virtues* (New York: Oxford University Press, 1983), Chapter I. Henceforth references to this work are paginated in the text.

He makes his case by pointing out, first, that a human life naturally divides into periods: infancy, childhood, adolescence, young adulthood, maturity, old age, and, second, that "such a division into 'times of life' tends to be accompanied, in most of us, by a sense of the greater importance or significance of certain times of life in comparison with others" (14). He maintains, for example, that "we have a definite tendency to discount youthful misfortune or success" (14), as well as the achievements and failure of senescence (18-21), as unimportant in comparison to those of mature adulthood. Slote contends that "Rawls, Sidgwick and others who have assumed the equal status of all times of life have not taken this sort of common judgment sufficiently into account" (15).

It is difficult to know, however, quite what to make of this empirical generalization about how we in fact view different periods of our lives (if it is a fact), without some degree of higher-level theorizing which Slote does not provide. Can Slote mean to say that the fact that we *do* view periods in our lives this way shows that it is *rational* so to view them? For surely Nagel *et al*. can concede that we may typically think of some periods of life as more important than others, without undermining their thesis that "*rationality implies* an impartial concern for all parts of our life. The mere difference of location in time, of something's being earlier or later, is not in itself *a rational ground for* having more or less regard for it."[28] Nagel *et al*. might explain our typical behavior in one of at least two ways. Either we may *irrationally* view a period of our life as more important than others, i.e., by exhibiting a time-preferential bias in that very judgment, or else we may view that period as more important than others, not because of its mere temporal location, but because of other concomitant factors contingently connected with its temporal location. Call these *contingency reasons*. In neither case would Slote be affirming a claim with which Rawls *et al*. would take issue.

Consider the first possibility. Doubtless the trials and tribulations of childhood and adolescence seem of momentous significance at the time, just as a burgeoning sense of self-confidence may lead the young adult to devalue the accumulated wisdom, insight, and tolerance of the aged. We may, indeed, typically overestimate the significance of temporally proximate states of affairs at every time of life, just because of their temporal proximity. But clearly it would be a mistake for Slote to maintain that from the temporal perspective of maturity, we take more seriously the successes and failures of maturity because of their temporal proximity to us. That would merely illustrate his claim that we have a pure time preference, not that it is rational to have it. Hence the perspective from which we evaluate a pure time preference as rational cannot be itself time preferential.

---

[28] Rawls, *op. cit.* Note 2, 293; italics added.

Moreover, it is not obvious that we do exhibit time preference in such cases as those Slote describes, even if he is right about the judgments we typically make. For (and this is the second possibility) there may be contingency reasons that explain why we take the achievements and failures of a particular period of life more seriously than others; for example, that they are the outcome of the full flowering of one's physical or mental capacities. If this occurred in human beings between the ages of two and five (say) as it does in dogs, we might favor a person's achievements and failures at that age instead. Thus these authors' answer to Slote might be that the reason we take a person's mature successes and failures more seriously than those of adolescence is because the former are her *mature* successes and failures, not because of the temporal location at which maturity occurs.

Slote appears to acknowledge this, when he says, "[W]hat we have so far defended is not 'pure' time preference, if by that one means the favouring, say, of earlier or nearer times of life as such. Rather, it is a preference for the goals and interests characteristic of certain states or periods of life rather than others, and these goals and interests are from a logical standpoint perhaps only contingently related to what comes earlier or later in time" (23). But the "logical standpoint" cannot be so casually dismissed, if Slote intends his readers to consider seriously the logical force of his thesis. If "the goals and interests characteristic of certain states or periods of life" may occur at any *time* of life (psychological lore has it, for instance, that intellectual maturity comes between eighteen and thirty for the mathematician, but between fifty and sixty for the historian), then Slote's empirical observations about our preference for those goals and interests are simply a *non sequitur* in relation to the philosophically compelling issue of pure time preference that Nagel *et al*. address.

When Slote then purports to turn his attention to this issue directly, he defends what he takes to be an even more radical thesis about the rationality of pure time preference. He claims that "even such pure time preference can be found (ironically) not in any favoring of the temporally nearer or earlier, but rather in a precisely opposite preference for what comes later in life" (23). But again the same difficulties arise: Does Slote mean to claim that, from the perspective of youth, we favor the experiences of old age? In our society this seems clearly false, but what would it show it if were true? For Nagel *et al.* it would show only the irrationality of youth. Or does Slote mean that *from no particular temporal perspective at all*, i.e., irrespective of our temporal location, we favor "what comes later in life?" This seems *prima facie* incoherent, since, when we make a judgment irrespective of our temporal location, we discount the temporal location relative to which the temporal location of any other event can be identified as "earlier" or "later" than it. Or, lastly, does Slote mean that, irrespective of our temporal location, we favor what comes later in our life as such, *whenever* that is, more highly than what comes earlier? In this

case, our judgment would be made, irrespective of our temporal location, about the significance of a certain kind of temporally characterized state of affairs, irrespective of its temporal location; call this a *temporally indeterminate* judgment. A temporally indeterminate judgment cannot provide evidence of any sort of time preference at all.

Slote neither addresses these questions nor furnishes the sustained conceptual analysis that might dispel them. Instead, he appeals to our intuitive responses to two cases: the individual who achieves great success only late in life and dies "while still 'in harness' and fully possessed of his powers" (23) versus the individual who achieves the same degree of success in her youth and then loses it permanently. His implicit question to us, then, is whether we would prefer to be what I shall call a *late bloomer* or an *early achiever*. Now suppose for the sake of argument that we would prefer to be a late bloomer, as Slote maintains. As before, there are at least two possible explanations for why we would, neither of which would controvert the irrationality of pure time preference. The first possibility is that we might be evaluating these two alternatives from the temporal perspective of already mature individuals who themselves enjoy or can anticipate great success only later in life, if at all, and so may be expressing a pure-time-preferential bias in our judgment of them. In this case, Slote's appeal to our intuitive response to these alternatives would beg the question. A second possibility is that our intuitive response might depend on contingency reasons. For instance, we may consider the late bloomer more fortunate because we envision his early failures as consoled by faith in his abilities, whereas we envision the early achiever's repeated failure to sustain her early success as exacerbated by her own deepening self-doubt and the pressure of others' disappointed expectations. But it might just as easily happen that the early achiever is an athlete, hence instead is both permitted and expected to rest on her laurels and to content herself thereafter by sportscasting or running a restaurant. Similarly, it might happen that the late bloomer is a mathematician, whose early failure to achieve success undermines his self-confidence and so poisons his life that his late, unexpected, and anomalous success provides insufficient psychological compensation for it. So the criteria by which we judge a person's life to be fortunate are not intrinsically related to the temporal location of its successes and failures.

Slote seems to reject the first possibility, i.e., that we ourselves are exhibiting a pure-time-preferential bias in preferring being a late bloomer to being an early achiever, when he says that he has "been concentrating on the sorts of judgments concerning individual good that are sometimes made when we consider the lives of others or stand back from our own lives and *attempt to view them in a detached way*" (28; italics added). Slote does not give a fuller characterization of this detached perspective. But we can infer, minimally, that he does not mean to maintain that we prefer a later period of

life merely from the temporal perspective of that period itself, or from a time immediately adjacent to it. Slote's argument must then be either that we prefer it from a temporally remote perspective, which is improbable given the age of his audience, or that we prefer it from no temporal perspective at all, i.e., irrespective of our temporal location. I shall assume he means the latter. The claim would then be that, from the detached, atemporal perspective, from which we make judgments of rationality, we do not necessarily have equal regard for all parts of our life.

Now Slote may think we prefer a later period of life from the detached perspective for any number of reasons. Perhaps it is because we typically (and, as we have seen incorrectly) associate maturity with later life. This explanation would fail to address the rationality of pure time preference, for contingency reasons. Or because we have a sentimental fondness for the twilight years of a person's life, including our own, whenever and at whatever age those occur. This, too, would fail to identify the phenomenon as one of pure time preference, for temporal-indeterminacy reasons. Or perhaps because we each just in fact prefer absolutely, from the detached, atemporal perspective, a certain identifiable temporal location as such, situated toward the end of each of our own lifelines, at whatever particular age it is situated, regardless or what occurs at it and what we typically expect to occur at it.

Surely we have no such preference. But, even if we did, this could not possibly prove the *rationality* of *pure time* preference. For the detached perspective from which we might, as Slote seems to think, prefer any such temporal location precludes the identification of that location as earlier or later than the location from which we prefer it. Hence it is not a pure time preference. And we have already seen that, if it were a pure time preference, we could not show it to be rational from that temporal location. Slote's dilemma can now be stated more generally. The perspective of a particular pure time preference conceptually precludes the detached perspective from which any arguments for the rationality of pure time preference could be advanced. Hence no Anti-Rationalist appeal to intuitions and examples of pure time preference can succeed in showing its rationality. On the other hand, any more abstract, systematic argument to this effect from the detached perspective implicitly repudiates the pure-time-preferential perspective it purports to defend.

Thus Slote's and Williams' Anti-Rationalist methodology, of appealing primarily to undefended basic intuitions, examples, and common-sense empirical generalizations at the expense of a more abstract and systematic (dare I say "rationalistic?") analysis, is ultimately self-defeating. For where they adhere consistently to this methodology, their claims are unlikely to persuade the skeptical, whereas, when they abandon it, they often ends up presupposing the view they purport to reject. In Chapter XIII I show that this

dilemma holds for Annette Baier as well. It would seem that Slote's dilemma is of a sort any consistent Humean Anti-Rationalist will find difficult to avoid.

3.2.3. The Impersonal Point of View

Next consider the possibility that an agent adheres to an impersonal point of view without adhering to universalistic moral principles. Recall that, on Williams' thesis, we can be detached from our ground projects in two ways: We may be detached from our own feelings and central aspirations, and we may be detached from others, in that moral theory obscures the reality of our circumstances, other people, and our attachments to them, by providing us with "one thought too many." In both cases, we lack personal reference or connection to our projects, by regarding them from a perspective from which they are not essentially ours. Now Williams would say that this just is the perspective of moral theory. But this is not necessarily so. To see this, consider the question of whether or not I should save my good friend Jeff first from some natural disaster; and suppose my moral theory sufficiently fine-tuned to yield the answer that I should, say, by the inclusion of a special-obligations-to-loved-ones clause. Williams' objection then would be that I am morally alienated from Jeff nevertheless, if I am motivated to save Jeff first because my moral theory prescribes it, rather than out of love for Jeff. Williams would say that my investment in this theory detaches me from my love for Jeff, since it is only in virtue of my theory that I am overridingly and unambivalently motivated to save him first. My impersonality is evinced by my primary attachment to my moral theory.

To see how very odd this objection is, consider the alternative it seems instead to recommend: I save Jeff first, but not because he is my good friend, nor because I love him, nor because I value and respect him especially as a person – since these are all descriptions that can enter into impartial moral prescriptions. Indeed, let us suppose that none of these descriptions are true of Jeff. Instead I save him first simply because he is who he is, namely *Jeff*, and for no other reason. Clearly this is absurd. I have salvaged Jeff's uniqueness and specificity, and the uniqueness and specificity of my relationship to him, at the expense of its intelligibility.

But the same objection could be made even if no such theory intervened between me and Jeff. For my desire to save Jeff first may also intervene between me and Jeff. In this case, the complaint would be that my investment in the satisfaction of my desires – especially, let us suppose, the altruistic and other-directed ones – takes precedence over my love for Jeff, since it is only in virtue of my unsatisfied desire to save him first that I am overridingly and unambivalently motivated to do so. This would be another example of the benevolent and other-directed but self-interested desires analyzed in Chapter VI.1.2. Here my impersonality, my lack of personal connectedness to Jeff, is evidenced not by my attachment to my moral theory, but rather by my

attachment to my own desires. My lack of personal connectedness to Jeff is maintained by my overabundant personal connectedness to my own desires and emotions.

In both cases, there *may* be some validity to these complaints. I may be, indeed, so enamoured of my moral theory and the fact that I subscribe to it that I really do regard other moral agents as nothing more than occasions for instantiating its precepts. But alternately, I may be so committed to the satisfaction of my other-directed desires that I regard other moral agents in a similarly superficial way, as mere occasions for exercising my beneficence. This is the stereotype of the "do-gooder," whose moral behavior somehow seems entirely self-aggrandizing, and neither elicits nor presupposes any attachment to the individuals her actions serve. In this case, too, the agent foregoes the depth and insight that accompanies attention to the specifics of who they are, for the sake of an essentially egocentric conception of reality. In the first case, the self is personally invested in a theoretical moral stance that degenerates into impersonality because it is used to deflect and disguise unmediated interpersonal contact. But in the second case, the self is invested in a *non*theoretical moral stance that also generates into impersonality because it deploys others as a means to the achievement of its personal moral ideal.

Thus I may subvert my personal connection with myself and others, not only by interposing a universalistic moral theory between us. I may accomplish this by interposing self-aggrandizing emotions and desires as well. To suppose that impersonality necessarily implies unemotionality would be mistaken. For in both of these cases, the very real problem to which Williams' thesis calls attention is not just personal detachment, but a deeper, more generalized pathological narcissism, of which the condition Williams characterizes as "moral alienation" is merely a contingent and localized symptom.

3.2.4. Narcissism

By *narcissism*, I mean that persisting state of the self characterized by an excessive preoccupation with one's self-image and image in the eyes of others, with one's personal flaws and assets and an unrealistic ideal of perfection against which they are measured, and with a self-oriented conception of one's relationship to others. Moreover, a self is narcissistic if it cannot tolerate others' independence of its expectations and requirements, nor appreciate their independence as intrinsically valuable. Finally, a self is narcissistic if these preoccupations effectively shield it against unmediated interpersonal

vulnerability or contact, and against the trauma of personal growth that frequently results.[29]

Among the manifestations of narcissism some psychologists have observed are envy and an inflated sense of one's own importance; the consequent devaluation of others and of their attentions to oneself; and the inability to empathize with others, to experience a sense of connectedness with them, and to form deep attachments to them. At the same time, narcissists are often highly dependent on others to buttress an extremely fragile and volatile self-esteem. They are beset by occasional eruptions of self-righteously sadistic anger and nightmares of self-contempt, both of which are recycled to fuel the grandiose belief that their ideals and aspirations are higher and purer than anyone else's. Consequently, narcissists are frequently smug, condescending, and seemingly remote as well.[30]

It is not difficult to understand how a narcissistic self might have a greater allegiance to its favored moral theory than to other people, or alternately, a greater allegiance to its altruistic self-image than to the individuals thereby served. Nor is it difficult to understand in what sense a narcissistic self might view others from a detached or impersonal perspective, and how its concern with the opinions of others might be accompanied by an inability to establish genuine and unmediated contact with them. That is, it is not difficult to see how what Williams describes as moral alienation might be a significant problem for a narcissistic self.

Let us speculate on the rationalized form such an attitude might take when the narcissistic self is confronted with evidence of its own narcissism, say, in the form of a complaint that one's aspiration to sainthood seems largely unaccompanied by any personal warmth. If narcissism functions as a defense against the intrusion of an undisguised other into the domain of the

---

[29] This characterization is based on the criteria described in *DSM III: Diagnostic and Statistical Manual of Mental Disorders*, Third Edition (Washington, D.C.: The American Psychiatric Association, 1980), 315-17.

[30] Dr. Otto Kernberg discussed these symptoms in a lecture at the University of California at San Francisco on May 2, 1984. Also see Kernberg's books, *Borderline Conditions and Pathological Narcissism* (New York: J. Aronson, 1975), Part II; and *Severe Personality Disorders* (New Haven: Yale University Press, 1984), Part III. For a related and insightful discussion that does not, however, use the concept of narcissism explicitly, see David Shapiro, *Autonomy and Rigid Character* (New York: Basic Books, 1979).

It is important to emphasize that the following discussion is intended to chart some of the connections between narcissism as I have defined it and the so-called impersonal point of view. It should *not* be taken to imply that actual proponents of Williams' thesis are narcissists, since of course it is the *philosophical* force of the thesis that has garnered it so many adherents. I am grateful to Jeffrey Evans for alerting me to this possible misreading of the argument.

self, a natural response to such a complaint would be to denigrate personal warmth as an invasion of privacy. Thus, for example, it might be argued that attempting to be familiar or cozy with everyone one meets is the worst form of moral inauthenticity of all; that it is irrational and self-indulgent to allow others to make importunate emotional demands on one; or self-defeating to allow them to take advantage of one's generosity and moral concern; that it is delusory or even self-destructive to "spread oneself too thin," "try to save the world," or be a martyr; and that the preservation of a private realm within which the universalistic demands of morality cease to apply is a necessary condition for having any sustained moral impact whatsoever.

These claims derive their persuasiveness from the recognition that few of us indeed can be morally effective when hanging from a crucifix. But deployed as narcissistic defenses, they rationalize the detachment of the self from the moral requirements of others, and the withdrawal of the self into a private domain in which those demands can be safely disregarded. By contrast with views that justify the need for privacy and personal fulfillment as a condition of greater moral compassion and commitment, the narcissist would claim that there is a certain realm in which the requirements of moral compassion and commitment simply fail to apply. That is, a sanctuary for the individual self is not justified as a necessary condition of sustaining and strengthening its moral ties to others, but instead as a sufficient condition of sustaining and strengthening the self to withstand them. Others, in this view, are regarded as intrusive or disruptive of the equilibrium of the self, or as exacting too great a demand on its resources rather than enriching them. The narcissistic self is distinguished, then, by the subordination of its moral commitments to its need for eminent domain.

A narcissistic self also might be expected to regard any universalistic moral theory of which it is an object as unsympathetic or distasteful. For by definition, a universalistic theory is impartial; that is, it refuses to accord privileged or exceptional status to the requirements of any self, including narcissistic ones. But we have already seen that one of the defining features of the narcissistic self is the arrogation to itself of value and importance that others are perceived to lack. From the viewpoint of the narcissistic self, the privileged status of its particular requirements, and its right to special treatment are justified by their special and superior value in its own eyes.

Moreover, from the perspective of a narcissistic self committed to the preservation of its internal boundaries against unmediated contact with others that threaten or disrupt it, the obligations to, for example, treat all human beings fairly, or not to be biased by one's personal preferences would naturally present themselves as particularly odious competitors to that commitment. For recall that another one of the defining features of such a self is the specifically self-orientation of its central desires, regardless of the content of those desires. I argued in Chapter II.2.3 and further in Chapter VI

that such an egocentric preoccupation with the condition of the self was a defining feature of the Humean self. We now can see the sense in which narcissism is built into the desire-based motivation of such a self.

The actual requirements of impartial morality disturb the integrity of the narcissistic self by threatening this orientation. For they demand, not merely a personal desire to conform to them, but rather an unmediated comprehension of and sympathy for the needs and requirements of others that, for a narcissistic self, are in direct competition with its own. We might expect, then, that the integration of personal needs and desires with the requirements of moral principle would be regarded by the narcissistic self as an abdication or sacrifice of selfhood, and rejected accordingly.

Not just moral alienation, but a more generalized social alienation is a predictable outcome for a narcissistic self. For if its primary concern with others is the nature of their relation to oneself, and if their behavior is invariably interpreted as evidence of this relation, then obviously, one's view of others will be mediated by this interpretation, and correspondingly detached from their independent reality. In this case, whether the terms of this interpretation are theoretical or affective is largely irrelevant. An agent who saves his good friend first in order to fulfill his moral obligations or satisfy his benevolent desires is psychologically and morally crippled, but not because of his moral theory. He is crippled because his preoccupation with his own rectitude overrides the dispositions and behavior that his moral theory prescribes.

Thus the real problem to which Williams' thesis importantly draws our attention does not lie with moral theory, or with transpersonal rationality more generally. For we have seen that one may adopt such a theory without assuming an impersonal point of view, and that one may assume this point of view without adopting a moral theory. Moral alienation is instead symptomatic of a more generally corruptive and debilitating pathology, namely narcissism, that bears no necessary relation to moral theory at all.

3.2.5. Self-Evaluation and Moral Paralysis Reconsidered

If the desiring half of the bipartite self thus has no more claim to psychological primacy than the reasoning half, then the problems of self-evaluation and moral paralysis do arise for Williams' version of the bipartite self, after all. The problem of self-evaluation is generated by the central role Williams accords to first-order desires, and takes a form similar to that described in Section 2.1, above. I argued in Chapter II.2.3 that from the perspective of one's first-order desires, all the apparent manifestations of the self are instrumental to their satisfaction; and that these desires are themselves, by definition, obfuscated even by our attempts to identify them. Such attempts then give rise to an infinite regress of instrumentally nested desires, the satisfaction of each n+l-level of which our n-level beliefs and

perceptions of the self are hypothesized to promote. Again a decisive commitment to the veracity and finality of any one such desire as providing the interpretive matrix for understanding the rest must be an irrational and unjustifiable act of faith. The problem of self-evaluation, then, is not ameliorated but exacerbated by the ascription of primacy to first-order desires in a bipartite conception of the self. For anything may be instrumental to the satisfaction of a desire, and any desire may be instrumental to the satisfaction of a further one.

The problem of moral paralysis also attacks Williams' version of the bipartite self, from two directions. First, it is inherent in the instrumental structure of first-order desires that no such desire can provide terminating criteria of evaluation of the motives of the self, no matter what their source. For any such motive, including desires, are subject to further scrutiny from the perspective of further desires to which they are presumed to be instrumental. Second, the bifurcation of the self into two parts implies that the problem of moral paralysis also arises in the same form for Williams as it did for Watson. For here, too, it is an open question with which part the self is identified on any particular occasion, and so an open question which part evaluates the other's motivational content. And this leaves open the possibility that, in situations of conflict, neither part may prevail. Again, the reality that we often do make rational and well-informed terminating judgments about our first-order motives, and are not ordinarily stricken with moral paralysis in situations requiring quick responses suggests that neither Watson's nor Williams' version of the bipartite self is adequate to the psychological facts.

Clearly, the problems of self-evaluation and moral paralysis can be generated by any multipartite conception of the self. Just as clearly, those problems are also generated by a unipartite conception of the self as structured and motivated by desire alone. The challenge is then to articulate an alternative unipartite conception of the self that both circumvents these problems and also respects the psychological data; and to consider whether the remaining in-house candidate, namely reason, might be adequate to furnish the basis for a unipartite conception of the self that successfully meets this challenge. In Volume II I argue that it is.

So Williams' sustained Anti-Rationalism does not succeed in supplanting reason with desire as the final arbiter of what ends we ought to adopt. While his analysis yields a more subtle conception of what a desire is, it answers the question of whether its object is rational or worthwhile by simply asserting its psychological centrality as a "ground project," to which the evaluations of moral or rational principle are irrelevant. But to this claim the appropriate response is the obverse of that made to Watson, and the same as that made to Frankfurt: Simply *asserting* that rationality is irrelevant to the assessment of final ends does not *make* it irrelevant. At this point in the discussion the Anti-

Rationalist may either table the discussion and back up her assertion with force; or she may provide a sound, rational justification for the irrelevance of rational justification to the evaluation of final ends. Neither alternative is likely to be rationally persuasive to the innocent bystander.

Frankfurt's and Williams' distinctive variants on the Humean conception of the self have in common a palpable impatience with the restrictions rationality and morality threaten to impose on objects of individual desire; and also a vehement insistence on the overriding worth of centrally definitive objects of individual desire, simply in virtue of their status as one's own. These two views, each in their way, exemplify and discursively defend the attitude Nagel analyzes as the subjective perspective, from the standpoint of which one sees and evaluates objects and states of affairs through the lens of subjective values that have the status of values solely in virtue of the agent's personal attachment to them. They also exemplify nicely the attitude I described in Chapter II.2.3 as "funnel vision." This attitude can be found, perhaps in its purest form, in very young children who have been indulged but not yet socialized to acknowledge others' rights and demands as fully legitimate. I offer a fuller analysis of this attitude in Volume II, Chapter V.6.1. This combination of devaluing rational justifiability and valorizing personal gratification is paradoxical in practitioners of a craft that valorizes rational justifiability and purports to ignore personal gratification as a criterion of professional evaluation. Neither Frankfurt nor Williams seem to appreciate the self-undermining implications of their responses to the problem of rational final ends.

This is unfortunate, since it is their allegiance to the Humean conception of the self that generates it. The problem of rational final ends is a problem only from the perspective of a conception of the self that ignores the fact that rational final ends can be distinguished from irrational ones, when confronted by the psychological fact that they do. An explanatory paradigm that requires us to deny or dismiss such basic facts about human nature is bound to generate problems that appear insoluble from within the perspective of that paradigm.

Chapter IX. The Problem of Moral Justification

In Chapter I.7.2.2, I argued that the problem of moral motivation that so preoccupied Nagel was one of three interconnected ones a viable conception of the self must solve. The second two were the problem of rational final ends and the problem of moral justification respectively, the latter being a special case of the former. With regard to rational final ends, I argued in Chapter VIII that the Humean assumption that no authoritative terminating criteria for rational final ends are possible generates further difficulties: of the infinite regress, of self-evaluation, and of moral paralysis.

Moral motivation and rational final ends are connected in the question of what kinds of mental events motivate action. If only desires do, and desires substantively understood are what Nagel calls "unmotivated desires" that simply assail us, then final ends – i.e. the objects we ultimately and noninstrumentally desire – simply assail us, too. No attempts at rational persuasion can alter them, any more than rational persuasion might alter a sudden craving for a jelly doughnut; and no celebration of their imperviousness to reason – under the rationales of decisiveness, wholeheartedness, or moral character – can validate those final ends that flout it. If, on the other hand, by "desires" we understand what Nagel calls "motivated desires," then these drop out of the equation; and rational deliberation-events alone may be able both to formulate final ends and to inspire us to achieve them. In this case our final ends would be corrigible by reason, and the transpersonally rational attempt to articulate and justify alternative ones would be a meaningful enterprise. If reason can identify final ends, then their moral justification is possible.

Assuming we can establish the power of reason to identify and set legitimate final ends, are there any beyond those we already have – for example, friendship or happiness – for it to set? Philosophers who engage in substantive theory-building in moral and political philosophy answer this question with an affirmative *fait accompli*. They fashion normative views that describe and defend detailed visions of the good life, just society, or right conduct that both offer alternatives to the final ends we already happen to have, and by implication criticize those we prereflectively strive to realize. That normative moral and political philosophers devote so much professional energy to advocating and defending the normative ethical visions in which they believe, even though these may fail to mirror the actual conditions under which they live and act, is *prima facie* evidence of the conative pull these alternatives may have on us.

Even avowedly Humean moral and political philosophers – those who purport to believe that final ends are not properly subject to rational criticism – exploit this position as a means of arguing for alternative sets of social

arrangements that will, they claim, enable citizens to maximize the promotion of their individual final ends, whatever these may be. It thus in effect defends this vision as though it were itself a final end that the reader, independently of her actual particular final ends, ought to adopt. Like all moral and political philosophers engaged in substantive theory-building, Humeans deploy the analytical reasoning skills of their trade in order to demonstrate the rationality and intrinsic value of these alternatives, regardless of the particular final ends any individual may have. But since the Humean conception implies that transpersonal rationality has a merely instrumental role in satisfying the agent's desires, Kantians who accept it are disadvantaged in their attempt to demonstrate that transpersonally rational final ends provide a viable alternative to the pursuit of desire-satisfaction itself.

The second question that then follows on the heels of the first is whether deliberation-events that identify new or alternative final ends it would be rational to try to achieve can also motivate action in the service of such ends, independent of desire-satisfaction. Certainly we can and do daydream about living differently, doing things differently, responding and interacting with one another differently than we have in the past. But can we actually be inspired to take concrete steps to realize such visions independent of the negative behavioral reinforcement against which we instinctively react? That is: can we *carry out a resolve* to take such steps, independent of the stick of painful past experience and the carrot of misplaced optimism? Again moral and political philosophers engaged in substantive theory-building presuppose that we can, even if this requires a sacrifice of comfort or desire-satisfaction on our parts. They assume we can be moved by the force of logic, by transpersonally rational considerations of justice, compassion, or perhaps even sheer disgust with our present level of moral turpitude, to override or subordinate self-interest – which often requires doing nothing and changing nothing – to what is best all things considered.

Moral philosophers thus try actively to inspire their audience to adopt their visions of the good life, and to guide action accordingly. Like the most shameless writers of self-help manuals, they use argument, exhortation, and example to communicate their visions of the good with the same vividness and conviction it has for them.[1] Thus by providing moral justifications of these visions that try rationally to persuade us of their value and importance to us, they aim to motivate our joint attempts to realize them. And sometimes, moral philosophers succeed in this endeavor, for good or for ill. Patanjali's

[1]Actually some of those self-help manuals can be pretty effective, provided that the author is both skilled and (very important) enthusiastic. Some recent volumes on time management are particularly worthy of the name "motivational literature." There are also some good ones on nutrition. Just because we cannot be rational dissuaded from a craving for a jelly doughnut does not imply that we cannot be rationally persuaded to desire broccoli sprouts.

influence on Gandhi and through him Bayard Rustin and Martin Luther King, Locke's influence on the American Revolution, Rousseau's on the French Revolution, Nietzsche's on World War II, Marx's on Communism, and Rawls' on "trickle-down" economics are only a few of the more prominent examples of the motivational influence of moral philosophy on concrete action.

Thus doing normative moral and political philosophy in order to develop and justify some substantive ethical vision of the final good is a quintessential example of invoking transpersonally rational justification in the service of moral motivation. In this way the problems of moral motivation, moral justification, and rational final ends converge in the practice of substantive moral and political philosophy itself. The very fact of normative philosophical practice as an enduring human enterprise of self-critique and self-development presupposes that these three problems can be resolved. This is the enterprise that the Humean conception of the self by implication rejects.

So baldly stated, this enterprise may seem impossibly ambitious and the Humean conception merely a realistic corrective. But the history of philosophy's influence on political events suggests otherwise. It is neither unrealistic nor dishonorable for a philosopher to hope that the resources he devotes to articulating his vision of the good will find their final expression in public social action by those who are convinced of its worth. If moral and political philosophers did not wish for this final outcome of their efforts, it is unclear why any such philosopher would be motivated to expend them. This chapter through to Chapter XII treats several late twentieth-century moral theorists who attempt to address the problem of moral justification within the constraints of the Humean conception.

But what does it mean to justify an alternative conception of the good? To justify a theory is to give convincing reasons for believing it – or for believing in it. Anglo-American Socratic metaethics models its justificatory methodology on those of logic, mathematics and the natural sciences. That is, it treats a normative moral or political theory as justified if it can be shown to be derivable from weak and intuitively acceptable foundational premises according to standard rules of inference. I agree with this methodology, and try in Volume II to further refine it. What counts as acceptable foundational premises differs from theory to theory, however, as does in what a derivation consists. In the late twentieth century three types of derivation emerged: Noncognitivism, Deductivism, and Instrumentalism; and most metaethicists in the Anglo-American analytic tradition continue to subscribe to one of the three. This chapter introduces all three and examines particular exemplars of the first two in depth.

Section 1 addresses *Noncognitivism*, a sophisticated variety of Humean Anti-Rationalism that attempts to derive substantive moral principles from noncognitive premises about attitudes and emotions rather than beliefs, assumptions or actions. However, the rationality that Noncognitivism rejects

as a criterion of argument and analysis it then resurrects as a criterion for evaluating attitudes and emotions: The idealization involved invokes reason as a criterion of the appropriateness of those attitudes and emotions through which we express our values, rather than as a criterion of deliberation, argument, or proof. It thus accords with the Humean conception in assigning a central motivational and valuational role to desire and emotion, and in relying on what is in the end the familiar model of means-end reasoning that characterizes the Humean model of rationality.

Working in the Noncognitivist tradition of Raz[2] and Gibbard[3], Elizabeth Anderson's pluralist, nonconsequentialist, rational attitude theory of value in particular offers a justificatory methodology that derives value judgments, and so alternative conceptions of the good, from foundational premises that accord *prima facie* significance to expressive attitudes but rely ultimately on a hypothetical process of interpersonal rational dialogue and deliberation similar to that proposed by Habermas, Dworkin, and particularly Rawls. However, Anderson's process of social rationality functions as a prescriptive criterion for directly evaluating expressive attitudes through which we communicate and reinforce shared values. This process therefore functions as a criterion of value relative to which substantive moral principles can be justified. I show how the process itself is biased in favor of conformist and socially conservative conceptions of the good; and conclude that it is therefore inclined to obstruct rather than promote the openness, flexibility, and social change that implementing a genuinely alternative conception of the good would presuppose.

Sections 2 and 3 examine *Deductivism*, the attempt to derive substantive moral principles directly from weak and widely shared metaethical premises through conceptual analysis and logical inference. Alan Gewirth's *Reason and Morality* is the most ambitious and comprehensive Deductivist attempt to date, within the Kantian tradition, to develop fundamental moral principles that both apply universally and imply specific and detailed solutions to normative controversies in applied ethics. Gewirth attempts to turn the Humean handicap into an advantage, by showing, essentially, that a substantive moral theory, that grounded in his Principle of Generic Consistency, is logically implied by the very concept of desire-satisfaction; and so that the latter end presupposes another, more fundamental moral one. He thereby means to derive this principle from the concept of action, by straightforwardly theoretically rational inference.

However, Gewirth's acceptance of the belief-desire model of motivation, according to which we necessarily have a "pro-attitude" toward the objects we

[2] Joseph Raz, *Practical Reason and Norms* (Oxford: Oxford University Press, 1990)
[3] Allan Gibbard, *Wise Choices, Apt Feelings: A Theory of Normative Judgment* (Cambridge, Mass.: Harvard University Press, 1990)

desire, generates problems for this project. From this "pro-attitude," Gewirth infers that agents necessarily value the purposes of their actions as goods (whether or not these purposes are in fact goods for them); and from this that they are logically committed to generic principles of freedom and wellbeing necessary for pursuing these perceived goods. He argues that such freedom and wellbeing are therefore rights not only to the agents themselves, but also to and for all agents. But agents do not need freedom and well-being in order to act purposively: A coerced or psychologically unstable action is still an action toward the purpose of which, according to the belief-desire model of motivation, the agent must have a pro-attitude in order to be motivated to perform it. Hence an agent's necessary pro-attitude toward the purpose of her action either does not imply that she values that action, or else that she may value coerced or self-destructive actions as well. Thus the tautological character of the belief-desire model of motivation enables Gewirth to infer too much from the resulting concept of action to justify a substantive moral theory. A different conception of action, as goal-directed behavior guided by an intentional principle, would avoid these objections without undermining his Deductivist justification of the Principle of Generic Consistency.

The problems generated by the Humean conception of the self are not confined to Kantians such as Gewirth who take for granted that there is no viable alternative to it. The Humean conception of the self subverts metaethical and normative moral projects even among those who explicitly adopt it. The problem of moral justification looms particularly large when the Humean conception of the self is deployed to provide an instrumental justification of a particular moral theory. By contrast with Deductivism, *Instrumentalism* is the view that, roughly, a moral theory is objectively justified if the actions or social arrangements it prescribes can be shown to be the most efficient means to an agent's final ends, whatever these may be. In Section 4 I introduce this view, and the wide variety of Humeans – Rawls, Brandt, Gauthier, and Harsanyi, as well as Hobbes, Locke, Bentham, Mill, and Sidgwick – who are each identifiable as Instrumentalists. Section 4 traces the general form of the dilemma for all such Instrumentalists: If the arrangements prescribed by the moral theory in question promote all final ends, then they are by hypothesis objectively justified, but do not exclude immoral final ends. If special motivational assumptions are added so as to exclude immoral final ends, they then do not promote all final ends, and so are not objectively justified. Hence modifying the Instrumentalist strategy in order to justify a moral theory either involves sacrificing objective validity, or else implies that the Humean requirement of instrumental rationality is doing no justificatory work. In either case, Instrumentalism can at best reaffirm the moral value of the ends we already have. But it can neither justify nor motivate the adoption of a moral theory that requires us to modify or sacrifice our ends, even for the sake of the common good. Chapter X and XI, following, apply this general

analysis to the particular theories of John Rawls and Richard Brandt respectively.

1. Anderson's Noncognitivism

1.1. Expressive States

To say that someone or something is *valuable*, on Anderson's account, is to say that it is rational, i.e. makes sense, for someone to value her or it. To *value* that person or thing intrinsically is to have a "complex of positive [or favorable] attitudes" toward her or it (2, 17, 124).[4] Favorable attitudes toward people or things are diverse: they include being inspired, attracted interested, pleased, awed, emotionally involved, attentive, or concerned (2). Anderson does not define what she means by an *attitude*, but she describes it first, as an expressive state, and second, as "partly constituted by norms that determine [its] proper objects"(3).

Let us leave aside this second clause for the moment and concentrate on the notion that an attitude is an expressive state of the agent. What kind of state is an expressive state? A brain state? A dispositional state? An emotional state? Or perhaps all of the above? What does an expressive state express? Emotions? Thoughts? Desires? Impulses? Perceptions? Or all of the above?

I shall describe a yes answer to these questions as the *inclusive conception* of an attitude. On the inclusive conception, an attitude can express emotions, thoughts, desires, impulses, perceptions, or any of the other myriad internal states – brain, dispositional, emotional, etc. – that constitute our mental life. And to say that an attitude, on this conception, expresses any such state is to say that the bare presence of this state necessarily manifests its particular contents in overt physical change – as, for example, inner turmoil might manifest itself in rapid breathing, dilated pupils, increased heart rate, and agitated movement of the limbs. So the inclusive conception of an attitude implies that for any internal state constitutive of our mental life, there is necessarily some observable, physical manifestation of it.

Now on Anderson's view there cannot be a necessary connection between any such state of the agent (whether brain state, dispositional state, or emotional state) and a *particular* physical manifestation of it. For that would conflict with her later attempt to distinguish appropriate from inappropriate expressions of an attitude (83, 129). If the connection between an internal state and a particular physical expression of it is a necessary one, there is no point in evaluating it as appropriate or inappropriate. That physical expression of it is not subject to reform.

---

[4] Elizabeth Anderson, *Value in Ethics and Economics* (Cambridge, Mass.: Harvard University Press, 1993). All page references to this volume are parenthecized in the text.

Similarly, Anderson's view cannot realistically imply that *any* such mental state – whether thought, desire, impulse, or perception – has *some* observable physical manifestation. For we learn to conceal and internalize our thoughts, desires, impulses and perceptions in the process of socialization; to keep our thoughts to ourselves, to suppress our desires, to refrain from acting on our impulses, to register our perceptions without reacting to them. This skill of controlling and internalizing our reactions is, in essence, what the process of socialization teaches; and, as Nietzsche observed, it is the origin of the interiority of our mental lives. But any such state that, for reasons of socialization, individual constitution, or personal control has no physically observable manifestation cannot be meaningfully described as an expressive state at all. So it cannot be true that, as the inclusive conception claims, all internal states constitutive of our mental lives are expressive states; nor, therefore, that all internal states constitutive of our mental lives conform to Anderson's conception of an attitude.

However, for later purposes Anderson will want to insist that anything that is an attitude is necessarily expressive. She will also want to insist that an attitude can be expressed appropriately or inappropriately. Since the inclusive conception of an attitude implies the rejection of both of these features, Anderson should reject the inclusive conception of an attitude.

By contrast, an *exclusive conception* of an attitude toward a person or thing might define it as a specifically emotional response, or disposition to so respond, to that person or thing, such that the emotion is caused in part by the agent's perceptions, thoughts and beliefs about the object of valuation. On the exclusive conception, an attitude is an expressive state in that there is a necessary connection between the agent's inner emotional state and *some* overt physical manifestation. However, it excludes any necessary connection between any such emotional state of the agent and a *particular* physical manifestation, since Anderson needs to be able to distinguish between the attitude itself and the appropriateness with which it is expressed.

For the same reason, the exclusive conception of an attitude excludes any necessary connection between the agent's inner emotional state and any intentional *action*, whether of execution or of omission. Thus the exclusive conception of an attitude excludes certain thoughts, desires, impulses, and perceptions unless they bear the right kind of causal connection to emotions. And it leaves open whether a particular attitude is expressed only in the most subtle and minimal overt physical changes, or in gross behavior or action of an appropriate or inappropriate kind. This account of an attitude appears to mitigate the objection I raised to Humean Anti-Rationalist views in the General Introduction, i.e. that they subvert in practice the enterprise of Socratic metaethics on which they rely in theory, by appealing to interpersonally inaccessible moral states to justify their moral judgments. The exclusive definition of an attitude mitigates the inaccessibility of such states

while suitably restricting their overt expressions. For these reasons, Anderson should accept the exclusive definition of an attitude.

Let us assume, provisionally, that she does. Then to value intrinsically a person or thing, is, first of all, to respond with positive emotions to one's perceptions and beliefs about her or it. However, not just any favorable emotional response to a person or thing counts as valuing it. Positive valuations must be "governed by distinct standards for perception, emotion, deliberation, desire, and conduct" (2,3). Parents who value – love – their children will feel not only proud when their children achieve, but alarmed when they are in danger, and disposed to rescue them. Parents who felt pride at their children's achievements but only indifference or regret rather than concern at their endangerment could not properly be said to value them at all. So to Anderson's account we can add that intrinsically valuing a person or thing requires not only a complex of favorable emotions and dispositions toward her or it, but also that this complex exhibit a certain internal consistency *determined by our concept of valuation itself.* Only a certain specific set of favorable responses, elicited under their appropriate conditions, counts as, e.g. loving or respecting or admiring a person or thing. That is, our valuation concept provides both a criterion for identifying the constellation of favorable attitudes constitutive of it, and also a standard of adequacy against which these responses can be measured.

On Anderson's account, valuing a person or thing in a particular way requires that this constellation of favorable attitudes – perception, emotion, deliberation, desires, and conduct "*express* and thereby communicate one's regard for the object's importance" (11). Here Anderson goes beyond the exclusive conception of an attitude I advocated above. That conception built in only the most minimal notion of expression, namely some overt physical manifestation of the agent's emotional state. The notion of expression Anderson invokes here builds in two further conditions: first, that an authentic valuation should "express [those] valuations in the world, ... embody them in some social reality ... actually establish the relationship to the object of one's concern which is implicit in one's attitudes toward it" (17). This seems right. An agent who is not moved to establish some such connection with the object she values either does not really or deeply feel what she claims to feel, or else her purported values have strictly armchair status. There is a natural causal link between emotion and behavior expressive of it in all cases, no less so in the case of that complex of emotions partly constitutive of valuation. If one really values something, one is disposed to act accordingly.

However, we can grant this much without requiring, as Anderson does, that the connection between valuing and action be a necessary one. The connection may be merely sufficient, such that a failure or inability to express one's valuation in action would not imply that it was not an authentic valuation after all.

1.2. Expressive Norms

Anderson's second requirement, however, is that in genuinely expressing one's valuations in action, one *thereby* communicates one's regard for the object's importance to some possible observer or listener. To do this requires that others can identify our behavior as appropriate, i.e. as meeting shared behavioral standards for expressing that valuation. This is the sense in which Anderson wants to claim that expressing valuations is governed by shared social norms relative to which others can recognize our behavior as expressing the valuations we intend to express by it. So in order to count as an authentic valuation, on Anderson's view, an agent must not only manifest overtly a positive valuational attitude. Indeed she also must not only express that attitude in action that establishes a relationship to the object she values. In addition, the action must be intelligible to others as an expression of her regard for that object. If the action through which the agent expresses her connection and regard for the object valued does not (or could not) communicate that regard to other agents, she does not qualify as authentically valuing that object at all.

This is a very strong claim. But there is no ambiguity in Anderson's formulation of it. She argues that "I am capable of valuing something in a particular way only in a social setting that upholds norms for that mode of valuation. ... To care about something in a distinctive way, one must participate in a social practice of valuation governed by norms for its sensible expression" (12). Earlier I suggested that intrinsically valuing a person or thing requires not only a complex of favorable emotions and dispositions toward her or it, but also that this complex exhibit a certain internal consistency determined by our concept of valuation itself. So, as we saw, loving one's children requires pride on some occasions, alarm and a disposition to rescue them on others. This suggestion was consistent with the exclusive definition of an attitude, since it did not require a necessary connection between these emotional and dispositional responses on the one hand, and action on the other. By contrast, we can now see that Anderson's account does require this, and more. Her idea is that one must conform one's action to the shared behavioral norms prescribing appropriate expression of a particular mode of valuation in order to be said to value something in that way at all.

The example she gives is that of honoring someone. Her claim is that if we do not physically do what counts socially as honoring her, e.g. treating her deferentially, applauding her or paying her obeisance under the appropriate circumstances, etc., we cannot be said truly to honor her. This seems right. But these are all actions whose connections to valuational mental states of the agent are contingent at best. Honoring may usually include valuational attitudes such as respect, admiration, perhaps affection, or esteem. But the

concept of honoring someone is not exhausted by the constellation of attitudes we take toward her, nor are any of those just named necessary conditions of it. It is much more closely linked to public performative rituals, and much more detached from any particular set of emotional states, in ways that other valuational attitudes are not. Honoring is in this way much more like promising or proclaiming than like appreciating or admiring.

To claim that one does not really value something unless one conforms to social norms for expressing that valuation seems much too strong, for it makes impossible inarticulate or concealed valuation, being at a loss to express the depth or intensity or particular quality of one's valuation of a person or thing, or of doing so awkwardly, or of trying to express one's emotional response and failing. To say that there is a natural link between authentic valuation and action is not thereby to say that there is a necessary link between them; nor to say that the action expresses that valuation successfully; nor that it does so in accordance with shared social norms for expressing that valuation.

A recent television commercial illustrates this point. A man takes a woman to a candlelight dinner at an expensive restaurant. The occasion is clearly a special one, through which he intends to express his feelings for her. Over drinks he gives her a small, wrapped gift, a jewel box seemingly intended for a ring. Her manner is expectant and loving. She opens the box to find a ceramic pin with the cartoon face of a clown on it. She looks up at him in shock and astonishment. He smiles uncomfortably. "Next time, better call 1-800-FLOWERS," the voiceover intones. The implication is clear that he has expressed his affection for her inappropriately. Perhaps he has. But perhaps he intended to express something else. Perhaps he meant to portray himself as a clown and symbolically give himself to her. Or to communicate that they should lighten up in this relationship and not lose their sense of humor just because they were getting involved. Or to suggest that he loves her for the madcap clown she really is rather than for the self-important façade she presents to the world; and to give her a means by which to advertise a more lighthearted self-image. The setting makes clear that he meant to show how much he values her, and also that he lacks the social resources for expressing in what exactly that valuation consists. Perhaps the culture contains no such resources, or perhaps he is just unschooled in the ways of love.

Thus as it stands, Anderson's analysis disses dorks, geeks, nerds, and dweebs. But even for the most highly socialized and sophisticated among us, sometimes there really are no words adequate to express our gratitude for another's support, nothing we can do to demonstrate the depth of our affection, no way to express our heartfelt appreciation – and simply saying this, or doing nothing, doesn't do the trick, either. This doesn't mean that we do not have those attitudes.

If conforming to shared norms for expressing our valuations were a necessary condition of valuing something, it is hard to see how new and original forms of social expression of these attitudes could arise, or what would motivate their creation, or how they could be recognized as original ways of expressing those attitudes. What of attitudes people in fact have that may not (yet) make sense to a society, such as the desire to rebel, detach, explore or innovate? What of societies that do have norms needed to adequately express its members' reflectively endorsed valuations, but none for valuations that are recognizably among the panoply of human valuations but not reflectively endorsed? And what of societies whose norms for expressing its members' valuations are themselves not reflectively endorsable?

Anderson agrees that "a social order can be criticized for failing to provide adequate normative vehicles for the expression of attitudes that have come to make sense to its members. ... If a society lacks the social norms needed to adequately express its members' reflectively endorsed valuations, the rational thing to do is to invent and institute such norms" (18). But if people can make reflectively endorsed valuations in the absence of adequate social norms for expressing them, then expressing those valuations in conformity with the norms cannot be a necessary condition for the existence of the reflectively endorsed valuations themselves.

1.3. Making Sense of Value

This brings us to Anderson's conception of a valuation's making sense; of its being reflectively endorsable. Just as Anderson equates having the constellation of favorable attitudes constitutive of a particular mode of valuation with expressing it in norm-governed action, she also equates rational valuation with expressing one's valuations through reflectively endorsable norm-governed action. Thus Anderson wants to distinguish between valuing something and that thing's being valuable. She improves on Mill's formulation by stipulating that something is *valuable* if it makes sense for someone to value it (91-2, 102, 124); and if it meets standards it makes sense for someone to value (114-5).

The concept of a thing's *making sense* is a central one for Anderson. Here she deploys two locutions. In some passages she speaks *interpretively*, of making sense *of* our attitudes. For example, she characterizes the quest for self-understanding as "an attempt to make sense of our own valuational responses to the world" (3); and the coziness of a bedroom as making sense of "[one's] feeling snug when [one] retire[s] there" (4). Later she suggests that "if either [of two very different and incommensurable ways of adequately expressing one's valuations of one's ends] makes adequate, but very different sense, of one's valuations, then reason permits the pursuit of either one" (63); and that "[o]ne can make sense of one's own attitudes only by taking up a

point of view from which others can also make sense of them" (95); and that "[t]o justify an evaluative claim is to appeal to reasons that make sense of particular attitudes toward the evaluated object" (97); and finally that "[j]ustification is concerned with making sense of our concerns and attitudes" (111).

In these passages Anderson treats some attitudes as a given. Making sense of them is then equivalent to finding the interpretation or explanation of given mental phenomena that makes them most comprehensible to oneself and others. On this reading, to justify one's valuations just is to explain them with reasons *why*. The interpretive locution has the advantage that it does not beg any questions about what valuations or attitudes any particular interlocutor might think one should have. It is a comparatively weak requirement on justification, in that it requires only that I understand your values, whether or not I share them. To complain that a valuation does not make sense is, on this reading, to complain not that it is personally unacceptable, but that it is unintelligible, i.e. that it violates certain basic conditions of conceptual coherence and consistency.

In other passages, however, she speaks *prescriptively*, of what it makes sense *for* someone *to* do. Were two friends to become enemies, she argues, "it would make sense for [one] to stop cherishing" an ugly bracelet given her by the other (19). Similarly, she says, it "makes sense for a person to value most [states of affairs] only because it makes sense for a person to care about the people, animals, communities, and things concerned with them" (20); and "what it makes sense to do now essentially depends on what one has done in the past" (34). Later she argues that "[i]f goods are not commensurable, then it does not make sense to maximize their values" (46). She defines a standard as "*authentic* if and only if ... it could make sense for a person to guide her responses by it. ..." and as *important* to a person "if it makes sense for her to care about it" (48). Similarly, she says that "[w]hich higher-order good it makes sense to use in justifying a person's choices depends on the context of decision ..." (54); that "it makes sense to value different good in different ways ...." (72), and that "the conditions that make states of affairs valuable are not other states of affairs, but the people animals, and things it makes sense to care directly about" (85).

In these passages, what it makes sense for someone to do is what there is reason *for* doing. To justify one's valuations is to demonstrate that the balance of reasons prescribes it. And to state that it makes sense for someone to do something is to state that the balance of reasons prescribes it. It is to advocate the doing of that thing. Thus it presupposes and expresses a set of values with which one's listener is assumed to agree. The prescriptive locution in this sense imposes a much stronger condition on justification than the interpretive one.

Anderson's explicit definition of making sense supports the prescriptive locution but not the interpretive one. To judge that one's valuations make sense, for Anderson, is "to judge that they would be endorsed" from a hypothetical, common point of view in which people can both achieve one another's valuations and also reflectively endorse them. The process of reaching this point of view as a desired endpoint is one in which "people interpret and justify their valuations by exchanging reasons for them" (3). Part of exchanging reasons for one's valuations is being "able to tell a story that makes sense of [an] ideal, that gives it some compelling point, that shows how the evaluative perspective it defines reveals defects, limitations, or insensitivities in the perspectives that reject these valuings" (92). Moreover, this process of justification is objective, on Anderson's account, if the participants in this dialogue can reach agreement or make progress when they adhere to the following norms of rational discourse:

> (A) they acknowledge the possibility of a permanent gap between their actual attitudes and rational ones;
>
> (B) they acknowledge the equal authority of others to offer criticisms and proposals;
>
> (C) no one competent to participate is excluded from the dialogue;
>
> (D) all apply reasons consistently to their own proposals and to others;
>
> (E) they aim for agreement or a common point of view;
>
> (F) they agree to work from mutually accepted reasons toward resolution of their differences;
>
> (G) the process contains methods for introducing new considerations as reasons and for criticizing what are currently taken to be reasons (93).

So one's valuations are valuable, i.e. make sense, if they would be endorsed by others who, through adhering to norms (A)-(G) constitutive of the process of rational justification, comparison and critique of one another's valuations, were to reach mutual agreement on their valuations. Thus this idealized process of justification is familiarly instrumental in form, with mutual agreement on valuations as the desired outcome. It is called for when people who have different values have some interest or need to reach agreement. It is possible, Anderson writes, when there is some overlap in the considerations each party accepts as counting for or against attitudes and judgments (93). And it is required as a necessary condition of making sense of oneself and one's own values (94-95).

This account of rational value does require that others share some values in common at the outset. It presupposes "a background of socially contingent and historically evolving social practices and conditions" (102), as well as common ground, minimally, in "shared intuitions or in curiosity, trust, and a

willingness to try alien practices" (105). It also presupposes that all participants agree either on what counts as a reason for or against something (norms (D) and (F)) or on how to introduce new considerations as reasons for or against things (norm (G)). Together these presuppositions constitute a quite substantial area of shared values. Anderson's claim is that unless one's valuations conform to the hypothetical point of view delineated by these norms, one cannot make sense *of* oneself or them at all.

1.4. Making Sense of Oneself

In this discussion so far I have been concerned to make room in Anderson's metaethics for valuational attitudes people can have but not express, or express but not in socially familiar or acceptable ways. I have wanted to insist on the existence of such *anomalous attitudes*, as I shall call them, even in the face of social ignorance or incomprehension of them. I will now want to insist on the rational value of some anomalous attitudes even though they do not meet some of the rationality conditions Anderson requires. I will want to show that such anomalous attitudes can be valuable, i.e. rational, even though they are not reflectively endorsable by other members of a social community governed by norms for their expression.

Consider someone whose valuational attitudes are marginal with respect to all the social communities in which he moves. Suppose, for example, that the background of social practices and conditions in which he was raised is at odds with the two in which he now lives and commutes; that these two are at odds with each other; and that the two in which he now lives and commutes have alienated him from the one in which he was raised. Also suppose that because of his outsider status with respect to all three cultures, his intuitions, perceptions and beliefs about the inhabitants of each are greatly at odds with the intuitions, perceptions and beliefs the inhabitants of each culture have about themselves, and similarly at odds with the intuitions etc. each culture has about the others. Suppose further that this outsider status has virtually sated his curiosity and willingness to try alien practices by requiring him, as a condition of his own adaptation and survival, to study and gain extensive familiarity with the mores of each culture – to become more knowledgeable about each culture, in fact, than any single inhabitant of any of the three cultures is about her own. Moreover, suppose his outsider status has brought upon him repeated and consistent social ostracism, rejection, and punishment, so that his ability to trust any member of any of the three cultures is virtually nonexistent.

It is not implausible that, as the result of his experience as an interloper in all three social communities, his conception of what counts as a reason in favor of certain basic matters might be equally at odds with others' conceptions, so that it would not be possible to apply certain reasons consistently both to his and to others' proposals (norm (D)). That certain

lifestyles were socially isolating, for example, might count as a reason against them for members of any of the three communities, but as a consideration in their favor for him; that certain activities would bring one face-to-face with one's own mortality might count as a reason against them for others but as a reason for them for him; that certain kinds of relationships would fill his life with connection to others might count as a reason against them for him but a reason for them for others; and so forth.

Similarly, it is not implausible that, given his experiences, he might not agree with members of any of the three social communities on how to introduce into the dialogue new considerations for or against things as reasons (norms (E) and (G)). By hypothesis he would be fully conversant with the practices members of all three cultures agreed on for doing this – something analogous, let us suppose, to following Robert's Rules of Order. But he might justifiably think these practices inadequate for introducing considerations that were radically unlike those members of these cultures were conditioned to recognize as reasons. He might think that precisely because of their social cohesion and conformity, there were certain sorts of quite important reasons that members of all three communities simply were not psychologically or socially equipped to consider; that they just wouldn't "get it." And he might think that only quite radical presentations of these considerations – in theatrical or otherwise dramatic symbolic form, perhaps, or in acts of self-immolation or antisocial destruction, might lead the light to dawn. Being unwilling or unable to perform such acts himself, he might conclude that there was no way for him to tell the story that made sense of his ideals, that would give it a compelling point or reveal the defects, limitations, or insensitivities in the perspectives of an audience of interlocutors whose experiences were so radically different from and limited relative to his own. Although he might fully understand their valuations, he might realistically conclude that there was no way for him to make his valuations intelligible to them.

For all of these reasons, his valuations would not be rationally endorsable by other participants in the rational dialogue Anderson describes, nor might he think it worth his while even to participate in it (norm (F)). But this would not imply that he was unable to make sense of his own values. First, it would not imply this for the interpretive locution. He would be able to explain his attitudes and values in the same terms I have just described, offering reasons *why* he values and disvalues as he does that enable us to understand his valuations even if we did not share them. Second, that his values were not rationally endorsable from Anderson's hypothetical common point of view would not imply that it would not make sense prescriptively for him to respond and act as he does. He could, by hypothesis, give realistic and well-grounded reasons *for* valuing solitude, silence, and confrontation with mortality, such that we would be compelled to recognize the rational integrity

of his perspective even if we did not share it. That these values would not make sense from the shared perspective of a community of participants engaged in rational dialogue of the kind Anderson describes does not imply that they do not make sense at all.

Third and most importantly, that these anomalous values were not thus rationally endorsable would not imply that our marginalized agent could not make sense of himself. Anderson argues that "one can make sense of one's own attitudes only by taking up a point of view from which others can also make sense of them. ... we can make sense of ourselves only by participating in practices of justification" (95). By contrast, this agent would make sense of his own attitudes from a perspective to which no one else had access, namely the perspective of having experienced the three disparate social communities in the unique and particular ways that he had. And he would be able to do this even though he had, by hypothesis, declined to participate in a social practice of justification of the kind Anderson advocates.

Now Anderson argues that "[m]aking sense of ourselves is not a matter of theorizing about an object whose properties we cannot affect. We make ourselves intelligible to ourselves by cultivating attitudes that make sense to us, by determining to act in accord with ideals we accept that have survived critical scrutiny" (91). She claims that part of the quest for self-understanding requires that when we recognize in ourselves attitudes that we cannot endorse from the hypothetical common point of view governed by shared social norms of discourse, we reform these attitudes "so that they make sense in the context of an enlarged self-understanding" (96). Thus self-understanding, on Anderson's view, requires active self-determination through the cultivation of attitudes that are reflectively endorsable from the hypothetical common point of view.

Again this is way too strong. It implies that we cannot make sense of what we cannot either endorse or improve; and so that those intractable and incorrigible parts of the self that are so necessary for bringing us face to face with our imperfections, our guilt, and our personal limitations must remain opaque or impenetrable to rational analysis. It also implies that we can easily improve what we cannot initially endorse, and I have yet to see an account of how this is supposed to work that does not degenerate into exhortations to bootstrap the triumph of the will over the flesh. Most people cannot even manage to stay on a low-cholesterol diet.

But my main concern is what it implies for the possibility of social and cultural change. Earlier I asked how new and original expressions of valuational attitudes could arise, if conforming those expressions to shared social norms were a necessary condition of their existence. The same question can be asked about new and original valuational attitudes themselves. Demographically mobile societies such as this one are constantly creating marginalized agents of the sort just described. Through upward mobility we

may move from our original class backgrounds to higher ones, and to different ethnic or cultural groups, through education and professional training; through downward mobility and economic contraction we may move in the reverse direction, and thereby into other new ethnic or lifestyle communities – perhaps even into homelessness or penal incarceration; through travel, new technologies, or contact with other cultures that create new possibilities for experience or lifestyle, we may find our most basic values or lifestyle preferences undergoing radical revision. Anderson conceives the relevant contrast along Marxist economic lines, as between individualistic and social conceptions of rational attitudes. But a society as marked by heterogeneous social values as this increasingly global one owes its plethora of anomalous attitudes at least as much to its ethnic and class diversity and mobility as it does to its capitalist economic structure.

Agents who undergo these social, economic and cultural shifts are regularly confronted with disparities between their own anomalous attitudes and those that are socially endorsed by the community at hand. Under these circumstances, one always faces the choice of to which source authoritative weight should be ascribed. Either one may conclude that one's own values are inappropriate, and take steps to reform them in accordance with the norms of the community; or one may conclude that the norms of the community are inappropriate, and take steps to reform them in accordance with one's own values. Those who are strongly identified with the norms of a particular community will incline to choose the former alternative; marginalized agents by definition have a greater capacity to choose the latter. Without this capacity, it is hard to see how social and cultural value change could occur.

When value change does occur, it does not require that one construct or even envision an alternative community that adopts and enacts the norms of rational dialogue Anderson describes, nor that one rely on such a hypothetical community to endorse and legitimate the anomalous attitudes one may know independently to be rational. A marginalized agent can recognize his anomalous values as rational if, to summarize briefly, (1) he can causally explain them by his experiences, (2) he can in turn cite these values as reasons for his behavior and attitudes, and (3) these values, and the experiences that form them, are internally coherent. Of course this does not imply that they are therefore morally acceptable to any actual or hypothetical community. Whether they are or not, what any actual or hypothetical community thinks about them is irrelevant to their rationality. In Volume II, Chapter VI.8 I show that this characteristic of rationality – its independence (or, if you like, its "individualism") – is crucial to understanding what motivates the whistle-blower to withstand the pressures and threats of his moral community for the sake of a higher good.

Therefore, social and cultural value change does not require that "if our lives are to be meaningful, then we must adopt a perspective informed by the

expressive theory as our global mode of deliberating about and justifying our actions, emotions, and attitudes" (83). Since rationality is itself a value, it is unclear why Anderson, after successfully developing a fully pluralistic theory of value, thinks she needs to ground it with monistic and global criteria of rationality. Social and cultural value change requires only marginalized agents whose anomalous values are sufficiently secure, independent of the community's, and well-grounded in their experience to furnish the distanced critical perspective from which the community's can be found to be lacking. That is the kind of agent for which I believe a genuinely pluralistic, rational attitude theory of value such as Anderson's can and must make room. The alternative is a set of values embedded in and reinforced by social conformity, convention and conservatism; values of a sort that only mire an actual moral community further in the habits of unreflective corruption with which it already is likely to be far too familiar.

2. Deductivism

Deductivism attempts to derive substantive moral principle via a conceptual analysis of the foundational premises that purport to generate it. In Chapter VII we encountered in the work of Kant and Nagel two examples of Deductivist metaethical strategies. We saw that just as Kant attempted to derive the moral law from the foundational concept of a free and rational being, similarly Nagel attempted to derive a principle of altruism from a self-conception of oneself as just one person among many – i.e. from impersonality and objectivity conjoined. In both cases, the argument proceeded by elaborating and explicating what was presumed to be contained in the concepts of, respectively, freedom and rationality in Kant's case; and being one person among many others in Nagel's. Deductivism, then, is a central technique of Anglo-American analytic philosophy applied specifically to the analysis of moral, political, or otherwise value-laden concepts.

However, Deductivism aspires to more than merely unpacking what is analytically implied by certain foundational moral concepts. It chooses which concepts to unpack with an eye to drawing forth from the analysis prescriptive moral principles whose rationale is to be found in the foundational concepts from which they are said to follow. This is not easy. On the one hand, the derivation must not have the tautologous form,

If P then P,

for fear of eliciting bored yawns. So, for example, it would not do for Kant to derive from the concept of freedom merely the principle that a free agent is not unfree, although that certainly would seem to follow by conceptual analysis of "freedom." On the other hand, however, the more interesting but suspect form,

If P then Q,

often elicits pained protests. So, for example, it is indeed very hard to see how Kant purports to get from the concept of freedom to the principle that a free agent always tells the literal truth in every situation, no matter what. Many do not think he does get there. To navigate successfully between these two extremes is, on the one hand, to derive moral principles that – like mathematical or logical proofs – really do follow by conceptual analysis from their premises; and, on the other, to derive from neutral or weak and widely accepted premises new and substantive principles that really do guide action in some identifiable direction.

On the face of it, it is very difficult to see how any metaethical justification can have it both ways. The problem is not that it is impossible to say anything interesting without going beyond the concepts with which one began. That is not true. We learn a lot by reading about what Aristotle thinks the concept of friendship entails, or what Kant thinks the concept of reason entails. The problem is rather that it is very difficult to draw forth principled prescriptions from conceptual descriptions – i.e. "oughts" from "ises." Yet this is essentially what Deductivism tries to do. Without lapsing into tautology, Deductivism attempts a proof that begins with premises that approach prescriptive neutrality as nearly as possible – so much the better to win your assent to them, whoever you may be; and from them attempts logically to derive conclusions that have rich prescriptive content which you are rationally – logically – required to accept; i.e. which are "necessitated by reason" (Kant), or "rationally inescapable" (Nagel).

Most metaethical Deductivists do not try to begin with premises that have no prescriptive content whatsoever. The concepts of freedom, reason, being one person among many others, objectivity, etc. are all value-laden to a certain degree, even if this amounts to no more than everyone agreeing that the states of affairs these concepts denote are worthwhile in some unspecified way. But a *Strict Deductivist* will rightly beware even of these relatively weak premises. For if we take it as a given that one cannot get out of a premise any more than one has put into it, injecting even this much prescriptive content into one's foundational premises threatens to tilt the derivation toward tautology, if its conclusions resemble them too closely in conceptual and prescriptive content (so, for example, from the premise of free agency, it would not be interesting to conclude that free agents should always chose freely). A Strict Deductivist will not be satisfied with tinkering with the content of the premises to the extent that the principles to be "derived" follow, by definition, as a foregone conclusion. Rather, a Strict Deductivist will set herself the more ambitious project of drawing forth a new and prescriptively

interesting result from weak, morally neutral and widely shared premises that were not previously thought to contain it.

However, avoiding tautology may tilt the derivation in the opposite direction, of fallacious inference: if one did not inject anything about freedom into one's foundational premises, it is very unlikely that one is going to get anything about freedom out of them, at least according to the standard rules of inference. So if the prescriptive moral principles one derives as conclusions have a great deal of highly satisfying content, whereas one's foundational premises are as nondescript and value-neutral as any premises can be, one has cause either for great exultation or – what is more likely – for serious concern. Either one has managed to achieve what no other philosopher in the Western tradition has; or else it is likely that something, somewhere in one's derivation has gone awry. The Strict Deductivist must take care, not only that she has not pumped into her premises the prescriptive content she wishes to derive, but also that she has not surreptitiously pumped into the steps of her derivation itself the prescriptive content she so scrupulously barred from her premises.

3. Gewirth's Deductivism

Alan Gewirth is a Strict Deductivist. In *Reason and Morality*,[5] he answers both of the questions with which this chapter opened,

> (1) Are there any final ends aside from desire-satisfaction it is rational to aspire to?
>
> (2) Can we be motivated independently of desire-satisfaction to accept and aspire to achieve such rational final ends?

in the affirmative. He both articulates a substantive vision of the good society as a final end, and, in order to motivate its acceptance, enlists a thorough and exhaustive analysis of possible objections and alternatives.

By attempting to derive a substantive moral theory from weak, value-neutral premises through straightforward conceptual analysis, Gewirth follows Nagel as a second and more recent contemporary philosopher committed to the Kantian tradition of Deductivism. What distinguishes Gewirth's approach from Kant's and Nagel's is the explicit nature of his commitment to the procedure of rational derivation through conceptual analysis. In the explicitness and rigor of this commitment, Gewirth takes his cue from John Rawls's similarly explicit and rigorous attempt, which I examine in Chapter X, to justify his moral theory as a derivation from the theory of rational choice. Gewirth matches Rawls's essentially Humean strategy with a Kantian one that is just as ambitious in scope and just as

[5](Chicago: University of Chicago Press, 1978). Henceforth all page references to this work will be parenthecized in the main text.

demanding in execution. The explicit nature of this commitment entitles us to inquire not only whether his justificatory strategy is successful, but also, more specifically, whether the rigorous standards of an explicit derivation have been met.

Gewirth's version of the Kantian strategy is the derivation of a normative foundational moral principle from a conceptual analysis, not of reason or of the self, but rather of action. This concept, at least on the face of it, genuinely is normatively neutral to a degree that those of reason or objectivity are not. The normative principle Gewirth derives is what he calls the "supreme principle of morality", the Principle of Generic Consistency (henceforth the PGC). The PGC commands us to act in accord with the generic rights to freedom and wellbeing of our recipients as well as of ourselves (135). The detailed elaboration of this principle, and its application to a variety of moral cases, describe his vision of the just society. This is the normative moral theory that occupies the second part of *Reason and Morality*, and that thereby answers the first question, above, affirmatively. The first part of his book comprises what purports to be a definitive justification for accepting the PGC as a guide both to social policy and to individual conduct, and so answers the second affirmatively.

3.1. Justification

By a "definitive justification," Gewirth means one that answers the "three central questions of moral philosophy" (3): First, there is the *authoritative* question: Why should I be moral? Second, there is the *distributive* question: Whose interests other than my own should be helped by my action? And third, there is the *substantive* question: Which interests are good?

Notice that the authoritative question calls for metaethical justification, not of any particular normative moral theory, but rather of one's commitment to some such moral theory or other. It assumes that we should be moral, and demands reasons why. The distributive and substantive questions, by contrast, call for substantive answers from some particular normative moral theory. In essence, they demand articulation of the basic prescriptions and values of the theory. Thus Gewirth's conception of a definitive justification requires a justification of moral commitment in general and an exposition of a particular set of normative moral premises. It leaves a gap between persuading us of the value of moral conduct in general on the one hand, and persuading us of the value of his particular account of moral conduct on the other. It does not explicitly require the justification of some such set of normative moral premises as themselves "rationally inescapable." However, Gewirth aims to close this gap by arguing that the overriding reasons for being moral in general necessarily commit us to the PGC in particular.

Gewirth also requires that the answers a definitive justification provides to the three central questions of morality be, first, *determinate*. That is, the

criteria of moral rightness these answers establish must have sufficient particular content so that the opposite content cannot also be derived from the justified principle (21, 164-5). This requirement concerns the content of the answers to the distributive and substantive questions, rather than to the authoritative one. It states that the justified principle cannot be so general in content that one could derive from it both an internally consistent set of moral prescriptions on action and also the negation of that set. An example of a principle that violates determinacy would be "Act to further your own self-interest," because it might generate prescriptions both to keep one's promises and also to break them, for the same situation. Another example would be the principle, "Do unto others as you would have them do unto you," because this might generate prescriptions both to render aid to the needy and also to withhold it under the same circumstances. The scope of application of these principles for particular kinds of situations would need to be specified in much greater detail in order to meet the determinacy requirement.[6]

By setting the determinacy requirement as a standard his own PGC must meet, Gewirth, like Rawls before him, sets himself the challenge of justifying a moral principle that specifically proscribes conduct a different principle might endorse, and so of making his theory palatable to those who might disagree with its particular prescriptions. By holding his own theory to the determinacy requirement, Gewirth signals his intentions both to take strong stands and to change minds. He rejects a traditional escape hatch for moral philosophers who justifiably want to claim as many converts as possible: of generalizing and thus weakening the particular prescriptions of their moral theory, or merely endorsing the familiar ones, in order to increase the breadth of its appeal. The harder task – Gewirth's task – is to specify the practical prescriptions of his moral theory as fully as possible, and also convince the unconverted of their worth.

Gewirth also requires, second, that a definitive justification be *conclusive*, i.e. that the criteria of morally right conduct implied by the PGC be both beyond rational challenge by competing moral theories and also categorically obligatory for all moral agents irrespective of particular circumstance (21, 23, 149-150). This extremely ambitious requirement concerns the answers to the distributive and substantive questions explicitly, and to the authoritative question indirectly. Gewirth thinks that particular standards of morally right conduct are beyond rational challenge by competing moral theories if they are implied by rational analysis itself; and that they are categorically obligatory for all moral agents if they are necessitated by some feature of moral agency that no moral agent can avoid. This is where Gewirth aims to close the gap

[6]See Henry S. Richardson, "Specifying Norms as a Way to Resolve Concrete Ethical Problems," *Philosophy and Public Affairs 19*, 4 (Fall 1990), 279-310 for a detailed analysis of what this would involve.

between the authoritative question on the one hand and the distributive and substantive questions on the other. He aims to demonstrate that the PGC is uniquely implied by a universal and necessary feature of moral agency, hence that the reasons why we should be moral are the same reasons we should follow the prescriptions derived from the PGC.

Thus Gewirth makes explicit the ambitions inherent in Nagel's project as well as Kant's. We have seen that in Kantian metaethics as traditionally conceived, moral justification is to be understood along the same lines as mathematical proof: the normative moral theory is a content-rich result to be generated from comparatively weak premises according to universally accepted rules of logic. Gewirth's metaethical thesis is that the rational analysis of action provides both necessary and sufficient conditions for answering the three central questions of moral philosophy. By "reason," he means the generally accepted canons of deductive and inductive logic and conceptual analysis (22-23). Gewirth's claim is that if we consider carefully what is essentially involved in action, we will see that the PGC follows from it as a matter of logical necessity. So he means to answer the three central questions of morality by arguing that we ought to be moral because morality is implied by action itself; and, more specifically, that *this* morality, as encapsulated in the PGC, is implied by action itself.

3.2. Derivation

Before turning to detailed scrutiny of Gewirth's reasoning, let us first examine the underlying traditional Kantian strategy itself, of deriving a content-rich normative theory from relatively weak premises through conceptual analysis. The strategy assumes, first, that normative moral truths – that is, the uniquely valid moral theory that applies universally and necessarily to all rational, human, and/or sentient agents – are as deeply embedded in our cognitive faculties as are logical truths. If P, and if P then Q, then Q, then if agents act, and if agents act then they are committed to the PGC, then they are committed to the PGC. Second, this strategy assumes that normative moral truths can be uncovered in much the same way as can logical truths, i.e. through a rational deductive procedure. Finally, it assumes that to do so is to insure for the moral truths thus uncovered the same cognitive and conative inevitability that logical truths seem to possess. Kant himself had good reasons for these assumptions. Does Gewirth? Suppose Gewirth's PGC can be shown to follow from the rational analysis of action. What are the implications for moral motivation, rational final ends, and moral justification?

Would we be rationally persuaded to adopt the vision of society implied by the PGC as a final end and obey its prescriptions, simply in virtue of learning that it was logically implied by the concept of action? Is this justification sufficient to inspire action on its behalf? How can a derivation of a moral principle from the concept of action give us any obligation to do

anything other than what we already do in virtue of acting? If the PGC is logically implied by the concept of action, then it follows tautologically from that concept. Then I necessarily follow it, just in virtue of acting. If the PGC requires me to do more than what I already do just by acting, then it *contains* more than the concept of action from which it is derived. In this case it seems that some extra assumptions must be appended as the argument progresses. Now Nagel appended to his account of our impersonal self-conception as one among many equally real individuals an analysis of reasons as timelessly, tenselessly, and objectively valid, in order to generate the requirement of altruism. This was unobjectionable because he did not claim to give a strictly logical derivation in the first place. By having chosen boldly to meet the Kantian challenge head-on, Gewirth does not have that luxury. If he really means to *derive* the PGC from the concept of action, then whatever is strictly implied by the concept of action itself should suffice to enjoin our obedience to it.

To see this, consider the consequences of my realizing that my being a teacher follows tautologically from my being a professor. There would be no such consequences. I would not behave any differently upon realizing that, in virtue of being a professor, I am also thereby a teacher. Then analogously, why should my learning that, in virtue of acting, I am committed to the PGC move me to do anything more than what I already do in virtue of acting? If my concept of action itself does not elicit from me obedience to the PGC, why should my discovery that that it implies the PGC do so?

Gewirth might respond by likening his conceptual analysis to a complex mathematical proof, in which the tautological relationship between axioms and results are not obvious, but rather must be exhibited through elaborate reasoning in an extended sequence of steps. Relative to this sequence, the conclusion is a surprise that affords us genuinely new information, even though it was implicit in the old. The question then would be how close the similarities between Gewirth's derivation and a mathematical one actually are. In a mathematical proof, a later step in the sequence is derived from an earlier one through the application of canonical rules of logic and theorems already proven. The challenge for Gewirth will be to not import into his analysis any additional unargued or controversial assumptions in order to derive the conclusion he wants. If he meets this challenge, then those of us with a deeply rooted commitment to accurate theoretical reasoning might be convinced to accept his derivation of the PCG from the concept of action. If he does not, then our acceptance or rejection of the PGC will turn on other, nonrational factors, such as its intuitive viability, familiarity, convenience, or appeal.

Thus it is important to emphasize that *the failure of the metaethical project of providing a foundational rational justification of a normative moral theory does not entail the failure, falsehood, or unacceptability of the theory itself, for Gewirth any*

more than it does for Nagel, Kant, Rawls, or indeed any of the other theorists considered in this project. We still might find the theory persuasive or inspiring, even we cannot justify rationally our doing so. If we cannot, this may mean either that no rational justification of the theory is possible, or merely that we have not yet found one. But in neither case would it follow that the theory was not true, or not the best we could do. The connection between truth and justification is not that close in any theoretical inquiry.[7]

Gewirth's solution to the motivational problem is to reason that if the PGC can be shown to be logically necessary, then its denial is self-contradictory. Then any agent with a minimal commitment to logical consistency has a conclusive reason for believing its prescriptions. But, Gewirth continues, to have a conclusive reason for believing that certain actions ought to be performed is thereby to have a conclusive reason for performing them oneself, "so that the principle's normative necessity, whereby its requirements for action cannot rightly be evaded, follows from its being logically necessary" (23). Gewirth's reasoning here echoes Nagel's argument that altruism is rationally inescapable because if one has objective reason to believe an end should be promoted then one has objective reason to promote the end oneself.

The first question, however, is whether believing that the PGC is entailed by the concept of action necessitates believing that an action that is in turn entailed by the PGC ought to be performed. If I believe that the concept of action entails the PGC, and that the PGC entails action A, then perhaps I ought to believe that the concept of action entails action A (we might most plausibly assume A to be some inescapably basic action such as minimally bestirring oneself). But merely deriving A from the concept of action does not in turn necessitate a belief that A ought to be performed. Ought I then do whatever is said to follow from the concept of action? If it *follows* from the concept of action, then presumably I already do. Do I have a choice about what implications of the concept of action to act on, i.e. what actions to perform? If these implications follow necessarily, as Gewirth claims, then presumably I do not; I perform them in virtue simply of acting. But if I did have such a choice, then it would be hard to see how any such logically entailed implications of actions could be obligatory rather than supererogatory. So either I have no choice but to perform such actions, or else I have no obligation to do so. Ought I do at least what minimally follows from the concept of action? For example, if the concept of action entails setting goals, ought I then set goals because I act? Of course not. If anything, it is the other way around: I ought to act because I set goals, and it is these goals that give me reason to act, not the concept of action that gives me reason to set

[7]See Paul Benacerraf, "Mathematical Truth," *The Journal of Philosophy LXX*, l9, November 8, l973.

goals. That A follows from the concept of action, then, is not sufficient to conclude that A ought to be performed. That A also follows from the PGC gives it a *prima facie* prescriptive legitimacy that its following from the concept of action alone does not.

But this legitimacy is only *prima facie*. For it is similarly difficult to see how my believing that the PGC entails A can necessitate my believing that A ought to be performed. In order for me to believe this, I must believe, not merely – or necessarily – that the PGC is entailed by the concept of action, but instead that there is reason to do what the PGC prescribes. That the PGC is entailed by the concept of action is no more reason to do what the PGC prescribes than that A is entailed by the concept of action is reason to do A. Gewirth should not try to argue that this entailment relation (assuming it is one) confers any particular obligation on agents. A less risky strategy would have been to argue that the PGC describes clearly what agents confusedly attempt to achieve whenever they act; and that the rational acknowledgement of the PGC as a clarified and elaborated concept of action entails certain obligations on the part of those who rationally acknowledge it as such.

The second question raised by Gewirth's solution to the motivational problem is whether believing that an action ought to be performed entails believing that one ought to perform it. Here are some of the actions I believe ought to be performed: Afghanistan ought to be rebuilt; batterers and child abusers ought to be sentenced to life in prison without parole; Lani Guinier ought to be appointed to the Supreme Court. But my believing that these actions ought to be performed does not necessitate any belief on my part that I ought to be the one to perform them.

Does the claim work for more generalized action-descriptions? Nagel's answer was that unless a tensed, subjective judgment from the personal point of view implied a tenseless, objective judgment with motivational content made from the impersonal one, practical solipsism would result. But does the arrow point in the other direction as well? Does an objective, tenseless judgment with motivational content made from the impersonal point of view imply a subjective, tensed judgment made from the personal one? No: I believe aid should be rendered to the needy, but (feeling pretty needy myself) not necessarily that I should be the one to render it.

The third question raised by Gewirth's solution is whether, even supposing I do have conclusive reason for believing I ought to perform some act A, such that I thereby have conclusive reason for performing it, I also have sufficient motivation for performing A. Is Gewirth an internalist or an externalist on this matter? This recurs to Nagel's attempt to rescue internalism from the Humeans, and the question is the same: Granted that I have reason to do A and accept these reasons as mine, is this acceptance sufficient to get me to actually do A? The range of answers also have not changed: If accepting, on good grounds, that I ought to do A, is an occurrent event, then

there is no reason on the face of it why this mental event cannot have the same conative power as any other occurrent mental event such as desiring or craving. If it is not, however, we await some explanation of how an abstract object such as a reason can function as though it were a material cause even though it in fact is not one.

3.3. Voluntariness

Gewirth's strategy is to attempt to derive the PGC from two of what he calls the *generic features* of action, namely its voluntariness and its purposiveness. His claim is that these features are *morally neutral* in that they "fi[t] all moralities rather than reflecting or deriving from any one normative moral position as against any other" (25). They are also *invariant* in that they "pertain generically to all actions" (25). Moreover, they demonstrate that action has a normative structure, i.e. that certain normative judgments are "logically implicit in all action" (26). Finally, Gewirth claims, these normative judgments themselves rationally imply the PGC. So actions *themselves*, on this account – not the *concept of* action, imply certain judgments; and these judgments, in turn, imply the PGC. Therefore actions themselves imply the PGC. I will want to call attention to each of these claims for the two generic features of action: their moral neutrality, their invariance across all actions, and, in Section 5 below, their capacity to generate normative judgments.

First consider voluntariness. Gewirth equates voluntariness with freedom by defining it as the control of behavior through unforced choice. So he wants to say that all actions have the property of being behavior controlled through unforced choice. But in the section following the above quotes, he says not that action is *in fact* behavior controlled through unforced choice, but rather that action *as envisaged by moral precepts* is behavior controlled by unforced choice:

> [T]he sense [of the word 'action'] relevant here is that which is the common object of all moral precepts as well as of many other practical precepts that set requirements for action. ... they have in common that the intention of the persons who set them forth is to guide, advise, or urge the persons to whom they are directed .... [I]t is assumed that the hearers can control their behavior through their unforced choice so as to try to achieve the prescribed ends or contents ... (26-27; also see 28, 30, 35).

Since there are many actions that fall outside the scope of moral and other practical precepts, this provision represents a significant restriction on the range of actions from which the PGC can be said to follow.

Assume for the sake of argument that individuals who issue moral and other precepts envisage the actions they prescribe to others as behavior controlled by unforced choice; and that this is what it means to act freely. Then individuals who issue moral and other precepts assume that those to whom these precepts are addressed can freely carry them out. But whether

this assumption is true or not is a separate question, and its answer may be independent of what any individual, including the agent addressed, thinks about it. The explanation of my carrying out a prescription you issue to me may be entirely unconnected to your belief, and indeed my belief, that I freely choose to do so. For example, I may have been socially conditioned by my upbringing to carry out reflexively just the sort of prescription you issued in just the tone of voice in which you issued it.

If the assumption that agents can freely carry out moral precepts is not true, this does not necessarily mean that an agent whose behavior conforms to such prescriptions does not perform bona fide actions. It may be simply that individuals who issue moral and other precepts are mistaken in their conception of what an action is. It may be that a person whose behavior is not controlled by unforced choice has nevertheless performed an action; and so that the individual who prescribed it has envisaged action incorrectly. Basing the derivation of the PGC on a conception of action as envisaged by those who issue moral precepts is risky if that conception turns out to be too optimistic and too demanding for the action-capacities human beings actually have.

So even if action as envisaged by moral and other precepts really is behavior controlled by unforced choice, this does not mean that is what action is in fact. If a behavior can be an action without being an action as envisaged by moral and other precepts, then voluntariness as Gewirth defines it is not invariant across all action, but instead only across those actions envisaged by moral and other precepts. Then the normative judgments such actions entail are not logically implicit in *all* actions, but only – perhaps – in those as they are envisaged by moral and other precepts. That is, they are logically implicit in a limited subclass of conceptual *representations* of actions, and not in actions *per se* at all. This narrows considerably the scope of actions that are supposed to entail the PGC; and to the extent that action as envisaged by moral and other precepts are envisaged incorrectly, undermines its justification.

Narrowing the scope of actions to those as they are envisaged by such precepts also calls into question Gewirth's claim that voluntariness – again, as Gewirth defines it – is morally neutral. It may be neutral among competing moral theories, but it is biased toward those which issue some moral precepts or other. As we saw in Chapter V.1.1, not all moral theories do this. Someone who does not already care what various moral precepts prescribe will not be swayed by what they assume in common about action. Someone who does already care, and wonders whether he rationally should, will regard what these precepts assume in common about action with an equally skeptical eye.

To this extent, whatever Gewirth can then derive from this conception of action will not answer the authoritative question of morality, of why in general we should be moral (2.1.(1), above). It will not try to supply reasons for being moral, but instead – like the authoritative question itself –

presuppose that we already are, at least to the extent of being positively disposed toward the acceptance of some moral precepts or other, without answering the question of whether it is rational to be so disposed. It will then try to convince us of why we should accept those particular precepts which are implied by the PGC. To say that Gewirth's conception of action as envisaged by moral precepts is not morally neutral is thus to say – at the very least – that it implicitly presupposes that the authoritative question already has been answered positively.

This necessitates some revision in our conception of Gewirth's justificatory strategy. Initially we understood Gewirth to intend to begin with the concept of action as a premise; from this concept to derive two generic features of it, from which in turn would follow certain logically implicit normative judgments – specifically, certain generic rights and right-claims; and from these to derive the PGC. This would follow the traditional Kantian Deductivist strategy, of deriving from weak and generally accepted premises substantive normative results through conceptual analysis. But upon closer examination, we see that Gewirth begins not with the concept of action, but instead with the concept of a moral precept. From this premise he derives a conception of action that he claims all such moral precepts imply; from this the two generic features, from these the normative judgments, and from these, finally, the PGC. That is, he begins with the formal concept of a moral precept in general, and finally derives a particular moral precept for which he claims universal application. The PGC, then, in fact is justified as following deductively from the concept of a moral precept. The concept of a moral precept is neutral among competing moral theories. But it is not morally neutral in its content, the way the concept of action is. Gewirth's particular conception of a moral precept contains particularly strong moral assumptions – choice, freedom, value, and power.

So far we have assumed that Gewirth's concept of action as behavior controlled by unforced choice is that envisaged by all moral precepts. We must now examine that assumption more closely. By the voluntariness of action, Gewirth means that the behavior does not occur from direct or indirect external physical or psychological compulsion, such as gusts of wind or terrorist threats; or from uncontrollable causes internal to the person, such as reflexes, ignorance or disease (31). When these mitigating causes are absent, Gewirth argues, the agent's unforced and informed choice

> is the necessary and sufficient condition of the behavior.... When there is such control, the person chooses on the basis of informed reasons he has for acting as he does.... The self, person, or agent to whom the choices belong may be viewed as an organized system of dispositions in which such informed reasons are coherently interrelated with other desires and choices. Insofar as a person's behavior derives from this system, it is the

> person who controls his behavior by his unforced choice, so that it is voluntary (31).

Voluntary action, then, is free action, i.e. action that expresses a disposition to act on the basis of informed reasons for choosing that action, such that the agent's unforced and informed choice of that action is the precipitating (not, as Gewirth would have it, necessary and sufficient) cause of the action.

By contrast, actions caused by forced choices such those performed under threat of bodily harm – from the gunman who demands either my money or my life, or the terrorist who demands either state secrets or the torture of my family, for example – "are not fully voluntary, hence not actions in the strict sense" (32). He admits that some behavior resulting from forced choice may be actions, if there is a moral precept prescribing or prohibiting it. For example, a moral theory might contain the precept that if the gunman will kill your friend unless you hand over your money, then you ought to hand over your money. But the fact that the act is done under duress leaves it with an "irreducibly involuntary" component. Although one chooses one alternative and controls one's behavior so as to carry it out, one's alternatives are set by another in such a way that failure to choose the lesser evil threatens the greater one. Since this component is not under the agent's control, Gewirth argues, it cannot be morally prescribed or prohibited. Because such behavior is to some extent compulsory and a response to threat,

> [t]his component, ... is not subject to [the agent's] control and hence is *not itself an object of moral or other practical precepts. For this reason, voluntary or free action in the full or strict sense excludes forced choice*, so that such choice is not included among the generic features of action *that provide the justificatory basis of the supreme principle of morality* (33; italics added).

The italicized passages make three points. First, if a behavior is not an object of moral "or other practical precepts," it is not, strictly speaking, an action "in the full or strict sense" on Gewirth's view. But how can this be? Can no behavior quality as a *bona fide* action unless someone does or could meaningfully prescribe it? What about whistling as I play in addition to as I work? Or doing the silly thing immediately after doing the right thing? Would Gewirth want to argue that whistling as I play or doing the silly thing are not *bona fide* actions because no one does or could meaningfully prescribe them? So that being silly never counts as a *bona fide* action? This will not do.

The italicized passages make two more points: second, if a behavior is not fully voluntary, it also is not an action; and third, it therefore is not among the generic features of action that justify the PGC. We will take these latter two claims in turn. Recall that Gewirth aimed to give a conclusive justification of the PGC, and that this required that it be beyond challenge by competing moral views. If Gewirth were successfully to derive the PGC from a broad conception of action shared at least by all such views (if not by all agents), it is not impossible that competing moral views would find this justification

difficult to challenge. But by defining action in such a way that forced choice actions are excluded, Gewirth immediately opens himself to such challenges from any moral view with a broader and weaker concept of action. Aristotle, for example, does not require that voluntary actions be free in Gewirth's sense. On the basis of a distinction between choosing an action and desiring it, Aristotle suggests that although such actions are performed under duress, they still may be voluntary because it is always within the agent's power to choose the less desirable alternative.[8] So Aristotle could challenge Gewirth's premises on the grounds that the latter defines voluntary action too narrowly.

Similarly, although moral precepts in Kant's moral philosophy enjoin us to perform certain actions as though we had voluntary control over our behavior, we are not excused from moral responsibility for behavior that indicates weakness of will, such as my bankrupting myself in the service of my empirical inclination to play slot machines. Actions whose maxims violate the categorical imperative are not for that reason excluded from the realm of action altogether. So Kant could challenge Gewirth's premises on the grounds that the latter defines action itself too narrowly. Gewirth's justification of the PGC cannot be conclusive if its underlying conception of action is so quickly susceptible to challenge by competing moral views that enjoy such broad support.

Certainly there is no unanimous agreement among action theorists about what an action is. But there are at least some generally acknowledged minimal conditions all actions must satisfy. The weakest of these is that an action must be *intentional*, i.e. goal-directed. This excludes such behavior as the kicking reflex, but might include such controversial cases as the blinking and coughing reflexes. It also might include other behavior that is goal-directed but not necessarily conscious. For this reason it seems too weak, and is at best a necessary but insufficient condition of action.

A stronger condition would be that an action must be not only goal-directed but also conscious. This seems insufficient because it fails to exclude the above cases of reflex behavior in which we are both conscious and behaving goal-directedly but are not conscious of the goal at which our behavior is directed. For this, formulating the condition as one of consciously goal-directed behavior, such that one is conscious of the goal, is only minimally better, since it does not decide borderline cases such as absentmindedly scratching an itch (you're conscious of wanting to relieve the itch, but only minimally aware, or unaware, of scratching it) or lighting up after having foresworn smoking (you're conscious – always – of wanting a cigarette, but light up out of reflex or habit, without attending to what you are doing at all).

---

[8] *Aristotle: Nicomachean Ethics*, trans. T. E. Irwin (Indianapolis: Hackett, 1985) Book III, 1110a16-17.

An even stronger and more controversial condition would be that an action must be at least partially caused by an *intention*, such that the goal is consciously held by the agent and the behavior is consciously aimed at that goal. But this can exclude such habitual but largely unconscious behavior as brushing one's teeth upon awakening, answering the telephone, or downshifting when making a turn. Attempts to redress this exclusion may add a dispositional condition requiring that the agent could and would identify consciously the relevant goals if asked, other things equal. But the dispositional condition obscures the location of the intention relative to the action: Is it a cause of the actual action that was performed, or merely a disposition that would be actualized counterfactually, i.e. only through questioning? I may sincerely explain that I meant to brush my teeth when you ask me why I stumbled to the bathroom upon awakening, without sincerely meaning anything when I actually do it.

The condition that an action must be not only intentional and the result of an intention, but must, in addition, be *deliberate*, i.e. such that the agent consciously performs a particular action in part because she perceives that action as the best available means to realize her goal, is much stronger than any of these. In fact it is much too strong, since it rules out waving rather than saying goodbye, ordering the trout rather than the swordfish for dinner, and so on. Gewirth's requirement that an action must be *voluntary*, and, even more strongly, controlled by *unforced choice*, is stronger yet, and so even more controversial. By now we are describing a conception of action so restricted in its application that only a few moral precepts may imply it.

So why does Gewirth insist on it, particularly since it restricts so narrowly the scope of actions from which the PGC can be derived and so the range of agents to whom the PGC is claimed to apply? Why – and this brings us to the third claim – can he not include forced choice under the rubric of action "in the strict sense" and argue instead that choice *simpliciter*, rather than unforced choice, is the defining mark of voluntariness? Even this would be an exceedingly strong condition to impose on what is supposed to be a general conception of action. But at least it would have intuitive plausibility, and it would buttress rather than undermine Gewirth's claims of universality and conclusiveness. In order to understand why Gewirth insists on unforced choice as a generic feature of action, we must turn to his analysis of its other generic feature.

3.4. Purposiveness

Purposiveness for Gewirth encompasses some of the criteria discussed above as characterizing a commonsense conception of action. By goal-directedness, Gewirth essentially means something between what I described above as the action's being the result of an *intention* and its being *deliberate*, i.e. that the agent consciously envisages some state of affairs to be achieved by the

action. The state of affairs envisaged may be the physical behavior itself, such as taking a walk; or it may be some causal consequence of the behavior, such as improving one's health. The purpose of the action is then the goal for the sake of which the agent acts. We saw in Chapter V.5 that since the description of the action is borrowed from the description of its purpose, it is this that may violate or conform to moral precepts.

Gewirth then goes on to characterize the goal or purpose of an action in the following terms:

> In this whole range, the agent's aims or intentions are wants or desires, so that in every action an agent acts more or less reflectively in accordance with his wants. These wants, however, need not be hedonic or inclinational; they may consist simply in the intentions with which actions are performed (38).

Here Gewirth identifies intentions with a certain kind of desire. He distinguishes between the intentional and the inclinational sense of desire (39). Whereas the second requires some hedonic element such as liking or taking pleasure in the object of desire, the first need not, and indeed may be pursued very reluctantly (38, 40). In the first sense, to want to do A just is to intend to do it (39). Nevertheless, to want in this intentional, nonhedonic sense to do A, "by the very fact that [an agent] aims to do the action" (40), Gewirth says, he has a pro-attitude toward it.

In equating intention with a certain kind of desire, Gewirth signals his metaethical allegiance to that Humean tradition that attempts to provide a reductive analysis of intention in terms of beliefs and desires.[9] Although Gewirth's distinction between the intentional and the inclinational sense of desire is reminiscent of Nagel's between motivated and unmotivated desires, for the most part they bisect each other. A desire to perform an unpleasant duty can be both intentional and motivated in the event that it is caused by deliberation. A desire to speak the truth at great personal cost can be both intentional and unmotivated in the event that it simply assails one. A desire to join the public library can be both inclinational and motivated in the event that it is caused by reflection on the wealth of satisfying books there to be borrowed. A desire for a slice of pie can be both inclinational and

[9]See, for example, Donald Davidson, "How is Weakness of the Will Possible?" in *Essays on Action and Events* (Oxford: Clarendon Press, 1980), 21-42. Davidson equates desire with the "inclination to act" (27) and intention with "value or desire" (31). Similarly, in "Intention and Akrasia," in Bruce Vermazen and Merrill B. Hintikka, Eds. *Essays on Davidson: Actions and Events* (Oxford: Clarendon Press, 1985) 51-74, Christopher Peacocke asserts in passing that "of course intentional action is always action on the strongest desire (in the motivational sense of strength) ..." (53). In both of these writers, and others in the Humean tradition, intention is taken to include beliefs and judgments about one's ability and/or likelihood of performing the action one intends. But these are taken to support the reductive analysis rather than furnish an alternative to it.

unmotivated in the event that one's anticipated pleasure in eating the pie simply assails one. Thus whereas Nagel's distinction addressed how desires are caused, Gewirth's addresses their content.

However, Nagel's concept of a motivated desire intersects with Gewirth's concept of an intentional desire in their scope of application, and this is where the lingering Humean sympathies of both surface. We saw that for Nagel, the explanation of a motivated desire to perform an action is identical to the explanation of the action itself, and that motivated desires seem to be logically but not materially necessary preconditions of action. For this reason, motivated desires could be both postulated to "explain" any action, and therefore eliminated as an explanatory variable. That is, that S had a motivated desire to perform act A was trivially true for any A, and hence contributed nothing to our understanding of A. – This was Nagel's strategy for honoring yet modifying the Humean, belief-desire model of motivation: to effectively eliminate it by acknowledging its ubiquity.

Gewirth has a more substantive plan for ubiquitous Humean desires. By identifying noninclinational wants with intentions, he goes further than most writers in the Humean tradition. Whereas they mean to explain intentions in terms of beliefs and desires, Gewirth, by contrast, explains intentions in terms of a certain kind of desire alone. By so doing, he effectively conflates the two rather different traditions in action theory discussed in Chapter V.6. As we saw there, the Kantian tradition defines actions as behavior motivated and guided by intentions.[10] An intention, in this tradition, must be conscious but need not be deliberate. The object of an intention is the goal or purpose of the action. But we have seen in Chapter VI.2 that one may seek this goal independently of any positive responses or favorable attitudes one may have toward it. So, for example, I may intend, and carry out my intention, to socialize with powerful colleagues toward whom I feel nothing but revulsion, and come to hate myself for my careerism as a result. Or I may intend, and carry out my intention, to fulfill an unpleasant moral obligation, even though it will net me no satisfaction whatsoever. On Gewirth's account, this is psychologically impossible, since "by the very fact that [the agent] aims to do the action he has a pro-attitude toward doing it and hence a positive or favorable interest in doing it" (40).

In the Kantian tradition, the intention that explains the action functions, first, as a precipitating cause of the action, i.e. as an occurrent mental event of setting oneself to aim at a goal;[11] and second, as a logically necessary condition of the action: I cannot do A without having intended to do A. As we

[10]Michael Bratman's "Two Faces of Intention," *Philosophical Review XLIII* (1984) and "Davidson's Theory of Intention," *ibid*. Vermazen and Hintikka, 13-26 provide careful defenses of a contemporary Kantian view.

[11]Peacocke (*ibid*. Vermazen and Hinktikka) denies that intention implies belief (69-70), but suggests that it does imply trying (68). Both of these theses seem to me mistaken.

have seen in Chapter V.6, the action itself depends on its antecedent intention for its very identity, and to perform a different action is by definition to have revised one's intention. Thus Kantians tend to identify intention with the will, and the close causal and conceptual connection between intention and action underwrites the claim that actions are essential expressions of will. But Kantians admit two different sources of motivation within a bipartite self, namely reason and inclination, and so two sources by which the will, and the formulation of intention, can be influenced.

This raises the possibility of a structural conflict between the different motivational components of the bipartite self: Reason may conflict with inclination. Thus the Kantian tradition explains the problem of weakness of will qua weakness of intention, i.e. of an intention the rationality of which is in question, and which is therefore susceptible to attack, dissolution or revision by contrary inclinations such as personal gratification or moral temptation. The Kantian tradition holds an agent responsible for akratic behavior because it assumes that the agent who performs it has revised her intention – and so the object of her will, and so her action – accordingly.

By contrast, the Humean tradition in action theory defines actions as motivated by desire and guided by beliefs about how to satisfy it. Alvin Goldman's theory of action was examined at length as a prominent example of this tradition in Chapter II.1.2. Achieving the object of the desire is the goal or purpose of the action. To seek this goal is by definition to have a favorable, "pro-attitude" toward it; and actually to achieve it is by definition to have desired to do so. So whereas one need not have a pro-attitude toward the object of an intention, one must have a pro-attitude toward the object of a desire, by definition of what a desire is. In this, the representational analysis of desire offered in Chapter II.2.1 fully accords with the Humean tradition.

A desire, in this tradition, must precipitate the action, but need not be an occurrent mental event. We have seen that, according to the Freudian variant on this tradition, neither the desire nor its concomitant beliefs must be conscious, much less deliberate; and that in the behaviorist variant favored by economists as well as some experimental psychologists, that an action achieves some goal is sufficient evidence for ascribing to the agent a desire for that goal, irrespective of the agent's reported mental state. So in this tradition, the explanatory connection between action and cause is retrospective rather than prospective. I may desire to do A, but my doing B instead implies that my desire to do B was stronger. If I did B in fact, the reasoning goes, I must have most desired to do B. Gewirth endorses this reasoning (40).

Thus the problem of weakness of will becomes a paradox on the traditional, monopartite Humean conception. First, since desires, not the will, motivate action, the action I actually take resolves any conflict among competing desires that may have preceded it. Second, even if I act in violation of some desire I flag as particularly central or meaningful, this only shows,

retrospectively, that I must have desired something different after all. Any internal, structural conflict I may experience over an action I have performed is a conflict, not between warring motivational components of the self; but rather between what I believe I desire and what I desire in fact. On the Humean analysis, beliefs are not motivational components at all, and irrespective of them, I always do what I most desire in fact. Since the desire to do what I did in fact is a ubiquitous feature of action, so is my pro-attitude toward it. How weakness of will can ever occur is then a mystery indeed.

Stephen Schiffer[12] analyzes weakness of will as the case in which an agent's first-order desire to ϕ is stronger than her first-order desire not to ϕ, her second-order desire not to act on her first-order desire to ϕ is stronger than her second-order desire to act on her first-order desire to ϕ, yet she ϕs anyway. Her first-order desire to ϕ is therefore stronger than her second-order desire not to act on her first-order desire to ϕ. That she ϕs in spite of her second-order desire not to act on her first-order desire to ϕ is what qualifies her action as weakness of will. I have already addressed some of the difficulties of this type of view in discussing Frankfurt in Chapter VIII.2. It can be added here that this is recognizable as a case of weakness of will only if the agent's first-order desire is irrational and her second-order desire rational; but neither Frankfurt nor Schiffer offer any assurance that it must be. For example, suppose ϕ is "to obtain adequate rest." Then in what sense am I suffering from weakness of will if I favor this desire over my second-order, reflective desire not to? Even Davidson's compelling example, in which I compulsively get up to brush my teeth even though this will make no difference to my dental health but will disturb my sleep, [13] preserves the rationality of the reflective desire I flout. Davidson may be right that the reflective desire or judgment I violate need not be moral. But it nevertheless must be *rational* in order to enter into an occurrence of weakness of will. Neither Frankfurt's nor Schiffer's analyses meet this requirement.

These two conceptions of action thus radically diverge. Whereas the Kantian tradition defines the consequent action with reference to its antecedent intention, the Humean tradition defines the antecedent desire with reference to its consequent action. So whereas the Kantian tradition implicitly assumes that actions depend for their identification on the independent and antecedent intentions that precipitate them, the Humean tradition implicitly assumes that desires depend for their identification on the independent and consequent actions they precipitate.

Marrying the two traditions as Gewirth tries to do engenders a very odd hybrid. Gewirth means only to equate intentions with a certain species of

---

[12] Stephen Schiffer, "A Paradox of Desire," *American Philosophical Quarterly 13* (1976), 195-203.

[13] See Davidson, *op. cit.* Note 9.

desire. But if actions are identified by intentions and desires identified by the actions taken to satisfy them, and if all actions are caused by desires, then all desires that we act to satisfy, not only the noninclinational ones, are identified ultimately by the intentions behind the actions they cause. But if desires were identified by the intentions behind the actions they caused, it would mean that identifying what I desire depended not, after all, on what I retrospectively did, but rather on what I prospectively intended to bring about: I could be said to desire O only if I intended to bring O about, and not merely if I in fact brought O about. This concept of desire would conflict not only with the theory of revealed preference out of which the Humean model of motivation has gotten so much mileage; but, even worse, with the underlying Freudian variant on which the theory of revealed preference – and so much else in social science explanation – depends. Thus I could have no desires unrecruited into my agenda for future action; no behaviorally manifested desires in which my own behavior first instructed me; no frivolous or innocuous or irrelevant desires; no subliminal or fantasy desires whose existence surprised me after the fact of their satisfaction. Their hedonic buzz would be a mere side-effect of realizing my prior intention to satisfy them.

This would be to regard each of my desires with a degree of seriousness not all of them deserve; and – more importantly, to abdicate the central tenet of the Humean belief-desire model that insures its universality. Gewirth might reject the Humean conception of desire on which they are based, according to which desires are identified by the actions they purportedly cause. But without this conception of desire, it is not open to him to declare that every intention – and so every action – implies a pro-attitude toward its purpose. The cost of circumscribing Humean desires by Kantian intentions is the ubiquity of those desires.Desires cannot depend for their identity on intentions because I can identify many of my desires independently of any intention to satisfy them, and can satisfy many of my desires without having intended to.

Consider how Gewirth's equation of certain wants with intentions then functions. The inference that we always have a pro-attitude toward the purposes of our actions is invoked to support Gewirth's later argument that we necessarily value our purposes as goods:

> It is important to have seen the connection presented above between purposiveness and wants or desires. For from this connection stems the fact that the agent necessarily regards his purposes as good, and hence makes an implicit value judgment about them; and from this, in turn, there necessarily follow other judgments, both evaluative and deontic, that finally entail the supreme principle of morality as a principle that every agent is logically committed to accept (41).

Now suppose we accept Gewirth's analysis of action. What follows from it? One of its implications is that if we have a pro-attitude towards *all* of our

purposes, whatever they are, then we must have a pro-attitude towards the purposes of coerced choice behavior such as handing over our money to the gunman, committing treason in order to prevent the torture of our family, and so forth.

Now we are in a better position to see why Gewirth was so concerned to deny that coerced choice behavior is "action in the strict sense." If coerced choice behavior is not really an action but instead merely a species of intentional behavior, then Gewirth's analysis of *action* will require only that we have a pro-attitude towards *unforced choices*. This requirement remains controversial, since surely we can voluntarily perform actions we find distasteful in every respect; but not as much so as the thesis that we have a pro-attitude toward actions we have been coerced into performing. Gewirth's denial that coerced choice behavior is *bona fide* action functions to circumvent this implication.

Unfortunately, eliminating coerced choice behavior from the purview of action does not eliminate it from the purview of pro-attitudes. Even if coerced choices are not actions, Gewirth's analysis of action still implies that we have a pro-attitude toward them, because on his view we necessarily have a pro-attitude toward all of our purposes, whether coerced or not. Gewirth could avoid this implication only by denying that coerced choice behavior is goal-directed, which is implausible.

So we are left with two alternatives. Either an agent's pro-attitude toward an action does not necessarily imply that she values it – since she has such an attitude toward coerced choices as well; or else it implies that she values coerced choices just as much as "action[s] in the strict sense." Gewirth's attempt to ground his Deductivist project in a Humean model of motivation has the unhappy consequence that obeying the PGC is of no more – or less – value to an agent than handing over her money to the thief who has a gun at her back.

3.5. Dialectical Necessity

Finally we turn to Gewirth's third thesis about the two generic features of action, namely their capacity to generate normative judgments. Here Gewirth introduces what he calls the *dialectically necessary method*. The basic reasoning behind the dialectically necessary method is that since thought is expressed in language and actions presuppose (at least to some extent) thought, tacit linguistic judgments can be ascribed to agents who act:

> [T]o the extent to which such practical thinking is attributable, and to some extent necessarily attributable, to the agent who performs actions as analyzed above,[14] to the same extent linguistic expressions or judgments are also attributable to him. This does not mean that he necessarily

[14]Note the qualification on action here.

> speaks aloud or mutters to himself vocally, but rather that in acting and thinking as he does the agent uses or makes judgments that can be expressed in words (42).

Gewirth's reasoning, then, is that in acting, the agent necessarily thinks; and in thinking, the agent necessarily makes implicit judgments we can ascribe to him.

This reasoning needs to be examined very closely. First, it is not obvious that I necessarily think on any level about what I am doing. Sometimes I do, sometimes I do not. I can act unselfconsciously just in case I act intentionally and achieve a goal I would confirm if someone were to ask me, but of which at the time I have no conscious conception. Driving a car or dancing or playing a musical instrument, among many other examples, can have this unselfconscious quality, and they, too, might be objects of moral precepts under given circumstances. Second, do I necessarily make judgments about what I am thinking about? Only if I think only and always in categorical declaratives, which most human agents do not. An agent can think about S, have S on her mind, without ascribing any predicate P to it; I defend this claim at length in Volume II, Chapter II.2, and rely here on its intuitive plausibility. Therefore an agent can think about S without making a judgment about it. So even if it were true, which it is not, that I always thought about my actions, it would not be necessarily true that *any* particular judgments would be ascribable to me in virtue of them.

Now Gewirth answers this objection by stipulating that the linguistic expressions ascribed to agents refer to dispositions to describe retrospectively what they did, if asked. But I may have no coherent answer, if asked, if my thoughts were not propositional in form; or if, since I really did not think about my actions at all, I have no retrospective speculations to offer. When asked, for example, what I meant to accomplish by placing the mushrooms on the exercise wheel and the hamsters in the salad bowl, I shrug my shoulders helplessly and respond forthrightly that I simply was not thinking about what I was doing. Thoughtless actions are still actions; and again they may count as such even under the restrictions Gewirth imposes. Moreover, retrospective interpretations of my own action may suffer the same handicaps as third-personal interpretations of that action. Having failed to retain in memory the details of my past, I may be unable to reconstruct accurately what my motives, goals and thoughts were at the moment of action. Certainly I can offer more or less plausible interpretations of what I must have had in mind. But these may have no more or less *prima facie* plausibility than their third-personal counterparts.

Gewirth's dispositional provision assumes that an agent can always produce a plausible verbal story of his actions on demand. He describes this as "practical thinking." The dialetically necessary method consists in rendering such thinking in equivalent, explicit linguistic expressions and

drawing out their logical implications. It is *necessary* because it examines statements the agent implicitly makes from within his standpoint that are "necessarily attributable to every agent because they derive from the generic features that constitute the necessary structure of action" (43-4).

> For example, the method does not proceed by saying merely that some person happens to say or think that X is good; rather, the method proceeds by saying that every agent necessarily says or thinks that X is good. The basis of this necessity is found in one or another aspect of the generic features of action and hence in the rational analysis of the concept of action. Thus, although the dialectically necessary method proceeds from within the standpoint of the agent, it also undertakes to ascertain what is necessarily involved in this standpoint. The statements the method attributes to the agent are set forth as necessary ones in that they reflect what is conceptually necessary to being an agent who voluntarily or freely acts for purposes he wants to attain (44).

But first, if the basis for claiming that every agent necessarily thinks X is good is the two generic features – voluntariness and purposiveness – of action in Gewirth's "strict sense," then these two generic features undermine rather than ground the claim. They circumscribe the range of action to those which have these two generic features, and so to those agents who perform this restricted class of actions. The population of agents excluded from this putatively universal claim would seem to be quite large, consisting, as it does, in all of those agents who almost always act under duress of one form or another: agents who, for example, are coerced by poverty into giving up their children for adoption, or into working at dangerous and ill-compensated jobs, or are coerced by threats of violence or death into remaining in abusive environments, or by fear of job loss or financial ruin into accepting brutality or sadism from colleagues. So even if it were true that every agent did in fact tacitly or dispositionally think about "what is conceptually necessary to being an agent who voluntarily or freely acts for purposes he wants to attain," it could not be true that, for any designated X, *every* agent *necessarily* thinks that X is good. Only those lucky enough to control their behavior through unforced choice would; and this would be a matter of contingency rather than necessity.

Second, is it true that every agent does tacitly think about "what is conceptually necessary to being an agent who voluntarily or freely acts for purposes he wants to attain"? In the above passage, Gewirth goes considerably beyond his account of practical thinking. The necessary statements he claims are attributable to the agent encompass not only expressions of the thoughts the agent is presumed actually to have in order to have performed the action she performed. They also encompass necessary conditions of free agency. However, no free agent need ever actually have thought about these conditions in order to fulfill them, nor need any such

agent demonstrate a disposition to make judgments about them when asked. Indeed, some such agents may exhibit a veritable aversion to thinking too much about what freedom entails, on the grounds that this tends to undermine their belief in it. Even if the two generic features of action were conceptually necessary to my agency, this would not imply that I necessarily thought about them; nor, therefore, that judgments about them were "within the standpoint of the agent," nor part of "what is necessarily involved in this standpoint."

The problem here is a conflation of what action from the third-personal, ascriptional standpoint implies with what action as envisaged – by moral or some other practical theory or interpretation, or by some agent who acts – implies; i.e. between behavior and action. From the third-personal, ascriptional standpoint, action is simply behavior. It is a physical event. Independently of its envisaging by its agent *or by a third party whose interpretation of it presupposes the ascription of judgment which is in question*, an action is an occurrence in the natural world just like other physical events such as earthquakes, rainy weather, or the growth of underbrush. Such physical events are not the kind of entity that can imply anything: What does an earthquake imply? What does a landslide imply? What does a potted plant imply? Similarly, the act of taking a walk, independently of its envisaged intention, consists in nothing more than an agent's behavior of repeatedly throwing his weight onto each foot alternately while pitching forward. Repeatedly throwing one's weight onto each foot alternately while pitching forward is not the kind of metaphysical entity that can imply anything.

Implication is an intensional, conceptual relation that can hold between propositions, statements, concepts, or terms. So a third-personal, ascriptional *interpretation* of behavior, or a *judgment* about behavior that ascribes an intention to the agent and thereby identifies it as an action, may have certain implications. The first-personal *intention* that identifies behavior as action for its agent may also have certain implications. So, for example, it follows from your intention to take a walk that you intend repeatedly to throw your weight onto each foot alternately while pitching forward. And in interpreting the behavior as action, a third-personal observer might *infer* something from it *considered as evidence* for or against some intention, proposition or theory. But an evidential relation is, too, an intensional conceptual relation between a theory and a physical state of affairs intensionally regarded as datum relative to it.

Thus an agent's physical behavior has implications only under some intensional interpretation, vision, intention, judgment, or description, and not otherwise. Under some such intensional interpretation etc., this physical behavior counts as action. Under others it may not. Under those in which it does, the interpretation in part comprises the judgments one ascribes to the agent. Judgments may have implications. But if one does not accept the

interpretation etc., one is not committed to the implications. So, for example, my behavior of putting myself through law school does not imply anything about "what is conceptually necessary to being an agent who voluntarily or freely acts for purposes he wants to attain," unless I explicitly conceive my behavior in those terms, or would if questioned about it, or a third person correctly interprets it in those terms. In the latter case, that person's interpretation of my behavior would determine the judgments ascribed to me, and so their implications. But the particular judgments that person ascribes to me also might be incorrect. There is nothing in my mere behavior that necessitates them.

However, Gewirth's argument depends on their necessity. He reasons that all agents "implicitly judge" their purposes to be good, and therefore that they have rights to freedom and well-being; that all agents who have purposes must claim these rights (i.e. must judge similarly); therefore every agent is logically committed to the generalization that all agents who have purposes have rights to freedom and well being (48). I show in Chapter X that this line of argument is, in outline, the reverse of that which Rawls defends. Whereas Rawls begins in the original position with free and equal agents who are claimed therefore to value their goals and therefore to possess self-respect, Gewirth begins with the definition of goals as valued by their agents; and on the basis of this premise argues for such agents' right to freedom. So an agent is logically committed to the PGC only if she "implicitly judges" her purposes to be good. But we have already seen in Chapters VI and VIII.3.2.2.2 that it is not a matter of necessity that an agent judge her purposes in this way; and in Section 3, above, that even if she did this would not, on Gewirth's view, suffice to qualify them as actions.

Were Gewirth's dialectically necessary method without these troubling complications, it would enable him to ascribe to the agent certain judgments about actions that the externalist – and Nagel – seemed to deny were motivationally effective. These judgments as particular mental events would be, on Gewirth's thesis, necessary logical implications of acting that were internal to the agent's personal point of view. As mental events, these judgments would be internal, subjective, and personal. As logically necessary implications of action, they would be external, objective, and impersonal. They thereby would furnish a psychologically integrated alternative to the practical solipsism that threatened Nagel's account. The internality of such judgments would tackle the problem of moral motivation, while their objectivity would tackle that of moral justification, and so that of rational final ends. Thus the potential importance of Gewirth's dialectically necessary method to satisfying his criterion of conclusiveness is clear. So are the problems that stand in its way.

3.6. Generic Goods

Gewirth believes that certain specific judgments are ascribable to agents in virtue of their acting, namely judgments about the goals they set themselves to realize:

> purposive action is conative and dynamic in that the agent tries by his action to bring about certain results or consummations that he wants ... to attain. ... He regards this goal as worth aiming at or pursuing; for if he did not so regard it he would not unforcedly choose to move from quiescence or nonaction to action with a view to achieving the goal (48-49).

Gewirth's argument here is that the very fact that an agent is motivated to pursue a goal demonstrates that he regards it as worthwhile, as valuable. But we have already seen in Chapter VIII. 3.2.2.2 that this demonstrates no such evaluation. An agent may have contempt or distaste for his goals, and for himself for pursuing them; or an agent may regard a goal neutrally, and find it of interest that he has and pursues this goal without finding the goal itself of interest in the least. Gewirth assumes that motivation implies positive evaluation because he accepts the Humean dictum that only objects of desire can motivate action. I argue in Volume II, Chapters II and V that this is not true. But even if the Humean model of motivation were the right one, it would not imply an equation of desiring something with evaluating it positively, because desiring is a psychophysical event whereas evaluating and judging is an intellectual one. Gewirth continues:

> This conception of worth constitutes a valuing on the part of the agent; he regards the object of his action as having at least sufficient value to merit his action to attain it .... The primary ... basis of judging something to be good is precisely its connection with one's pro-attitude or positive interest or desire whereby one regards the object as worthy of pursuit. And since it is admittedly some desire, at least in the intentional sense of wanting, that provides one's purpose in action, it follows that an agent acts for a purpose that constitutes his reason for acting and that seems to him to be good on some criterion he implicitly accepts insofar as he has that purpose (49-50).

But desire, even in Gewirth's "intentional sense," is not the only source of purposes of action. Resolutions, choices, inferences, and external states of affairs such as events or other people are, in addition to intentions, further sources of goals an agent may adopt, none of which necessarily have even the most tenuous relation to desire.

Since, as we have seen, an agent need not value her purpose, she need not regard it as good. Nor need such a purpose constitute her reason for acting (although of course it might still constitute the explanatory reason why she acts). Nor, therefore, does such an agent therefore value the generic conditions of action necessary for achieving such purposes, namely freedom and well-

being (52, 53). On Gewirth's view, well-being includes the follow *generic goods*: *basic goods*, i.e. all necessary preconditions of acting; *nonsubtractive goods*, i.e. those whose use does not diminish an agent's level of purpose-fulfillment; and *additive goods*, i.e. those whose use increases an agent's level of purpose-fulfillment. But since, on Gewirth's view, we always have a pro-attitude towards whatever we do, it is hard to see how any action we might take could lower our level of purpose-fulfillment. And so it becomes a trivial truth, on Gewirth's view, that every action maintains or increases wellbeing (55), but not a truth at all that any agent necessarily values this. Since an agent need not value freedom and wellbeing, nothing follows from the concept of action about our mutual obligation to protect the freedom and wellbeing of agents. The PGC in particular does not follow from it. Thus does Gewirth's allegiance to the Humean conception lead his potentially powerful derivation to unravel.

4. Instrumentalism

Instrumentalism is a special case of Deductivism. We saw in Section 2 that Deductivism tries to derive normative moral principles from foundational premises via a conceptual analysis of those premises. Instrumentalism similarly tries to derive normative moral principles from foundational premises. In this case, the premises specifically concern what human beings are, what moves them, and at what they aim. The method of derivation is, again, conceptual analysis of those premises, i.e. of the properties thus ascribed to human beings. These combine to form a conception of the self that is the background theory from which empirical predictions are then derived about how, given those properties, human beings will behave and what they will choose in order to achieve their goals. Thus the procedure includes empirical conjecture and inductive as well as deductive inference. The outcome is an argument that justifies a set of normative moral principles as those which an agent would choose as optimally instrumental to his final ends.

The Humean conception of the self lends itself to an Instrumentalist strategy, because its conception of rationality is Instrumentalist in nature: The self is conceived as rationally coherent to the extent that theoretical reason calculates and schedules the satisfaction of as many of its desires as possible, with the minimum necessary costs. So we make sense of an agent's behavior by ascribing to her the desire to achieve the ends that she does in fact achieve, and the theoretically rational belief that, given the information and resources available to her, behaving as she did was the most efficient way to do so. The moral principles she is claimed to choose under certain given circumstances to govern her social relations are then argued to be the most efficient means to the achievement of her final ends, whatever these may be.

Instrumentalism is a *strategy* of moral justification because of the relation it bears to the Humean conception of the self. If you believe this conception to

be true of yourself and other human agents, then if you want to motivate other agents so conceived to accept your favored normative theory – or, for that matter, any suggestion of yours, you must demonstrate to them that what your theory enjoins them to do is in fact the most efficient thing for them to do, in order to achieve the desired ends they already have. And Instrumentalism is a strategy of moral *justification* because it attempts to persuade other agents that your suggested theory is objectively the right theory. In this way Instrumentalism is not merely a bit of practical reasoning that directs someone to perform certain actions in order to maximize the achievement of his given ends. It is more than that, because it attempts to demonstrate that reason directs *all* of us to perform those very same actions in order to achieve *any* of our given ends. This means not only that each of us has our own reasons to perform the very same actions that are in fact prescribed for others. The persuasive appeal of the theory in question is heightened to the extent that the Instrumentalist strategy can show its prescriptions to be instrumentally rational, not just to your final ends, but to anyone's. This fact about these prescriptions, if it is a fact, is supposed to provide you with a reason to conform to them that is independent of their instrumentality in promoting your particular ends.[15]

To the extent that Instrumentalism is successful in providing an objective justification of a normative moral theory – and we will see that it cannot be completely successful – it cannot provide a moral justification. But when we try to modify it so as to produce a specifically moral justification, we undermine its objectivity. In this case, we are forced to conclude that either an objective, Instrumentalist moral justification of a normative theory is foreclosed, or else the Humean notion of instrumental rationality is doing no justificatory work.

4.1. Instrumentalism and Objectivity

The motivation behind Instrumentalism as characterized above is not difficult to understand. The Cambridge Platonists failed to justify moral statements as referring to objective facts directly deducible from theoretical reason,[16] and the Moral Sense Theorists and Emotivists tried to demonstrate the implausibility of belief in any such facts.[17] Nevertheless, many of us

[15]Richard Brandt is particularly explicit about his use of this strategy in his *Theory of the Good and the Right* (Oxford: Oxford University Press, 1979); see particularly Chapter VIII. Brandt's view is discussed in Chapter XI, below.

[16]See, for example, the selections by Cudworth, Samuel Clarke, and Wollaston in D. D. Raphael, Ed. *The British Moralists 1650-1800*, Volume I (Oxford: The Clarendon Press, 1969).

[17]For the former, see Francis Hutcheson, "An Inquiry Concerning Moral Good and Evil," in D. D. Raphael, *ibid*. For the latter, see, for example, Charles Stevenson, *Ethics and Language* (New Haven: Yale University Press, 1944).

continue to believe that our deepest moral convictions have the same sort of claim to objective validity as our epistemological convictions (whatever sort that may turn out to be), to the extent that they are equally fundamental psychologically. It is natural to view what I have called the Socratic enterprise of analyzing and rationally evaluating theories as a natural extension of the prephilosophical impulse to question, criticize, and modify those convictions in light of evidence and argument. From this perspective, a convincing case has yet to be made for exempting moral beliefs and theories from these practices. While we may agree that moral truths cannot be deduced from reason or directly confirmed by the "furniture of the earth," many of us are less easily persuaded that, as the Emotivists claim, moral beliefs are not genuine beliefs at all. However, our awareness of the history of moral philosophy confronts us with the dilemma of what connection between our moral beliefs and the requirements of objectivity might be left to us to argue for.

The Humean conception of the self furnishes a substantive solution to this dilemma. Just as its motivational constituent supplies a strategy for motivating other agents, so conceived, to accept one's favored theory of what they should do, similarly its model of instrumental rationality supplies a connection between that theory and the requirements of objectivity. The Humean model of instrumental rationality accepts the traditional conception of fully informed, theoretically rational belief as objectively justified belief, and then assigns such belief an instrumental role in achieving the agent's desired outcomes. Action is then rational, hence objectively justified, to the extent that theoretical reason identifies it as similarly instrumental in producing that outcome.

The implicit reasoning can be reconstructed as analogical. One necessary requirement for viewing a scientific theory as objectively justified is that, oversimply, it accurately predict certain consequences of a set of given events. Similarly, the Humean might say, we may view an action as objectively justified to the extent that it is implied by a theory that accurately predicts consequences of so acting which the agent in fact wants to effect. So, for example, if a theory accurately predicts that keeping promises will enhance social stability, then I may justify my keeping promises on the grounds that I want to enhance social stability, and keeping promises enhances social stability. Here it is assumed that to the extent that I act in accordance with the theory's prescriptions for producing those desired consequences, my action is more likely to achieve those desired consequences in fact. Of course this assumption may be wrong, even if the theoretical reasoning that engenders the theory is correct and its predictions accurate. No further justification of the correct theoretical reasoning that generates the theory is needed, because correct theoretical reasoning about the facts itself constitutes the terminating criterion of theoretical rationality as asymptotic to objectivity. Hence an action

taken on the basis of correct theoretical reasoning about its predicted consequences receives the imprimatur of objective validity derivatively, in virtue of its instrumental connection with theoretical rationality.

Instrumentalism then extends this line of thought to the justification of normative moral theories, by attempting to demonstrate the objective validity of a moral theory as the most theoretically rational means to a wide range of unspecified final ends. Here one's favored theory plays the same role relative to the Humean model of instrumental rationality as does an instrumentally rational action. Just as an action is objectively justified by its expected instrumental success in achieving an agent's final ends, similarly, it is claimed, with the correct normative moral theory. A theory lays claim to objective validity if the actions or set of social arrangements it prescribes are the most instrumentally rational means to an agent's final ends, *whatever they may be.*

This last clause represents a more ambitious extension of the concept of objective validity just described. That concept connected theoretical reasoning with accurate prediction of objective events. But in the case of a scientific theory, we require some further, independent check on the accuracy and comprehensiveness of the theoretical reasoning by which the theory was constructed, in order to insure that the events it predicts are objective ones. In particular, the theory's predictions must be independently confirmable by other relevantly placed, disinterested observers under similar experimental conditions. Successful independent confirmation then elicits the intersubjective acceptance of the theory by such observers.

Again the application to moral theories is analogical.[18] Under comparable conditions, the Instrumentalist might claim, we each may be moved intersubjectively to accept a moral theory as theoretically rational and so objectively valid. That acting on the theory's prescriptions actually promotes the range of ends it is predicted to promote confirms the theory's theoretical rationality to each agent considering whether or not to accept it. This is objective evidence that the theory is in fact theoretically rational and hence objectively valid, and not just that it appears to be to some particular agent whose information and reasoning powers are limited. And that acting on the theory has predicted consequences that are desirable to other, relevantly placed agents who take an interest in the particular predicted consequences I happen to desire is evidence that the theory's theoretical rationality and objective validity do not depend on the particular ends I happen to have. To show that the actions or set of social arrangements a normative moral theory

[18]The analogy between inductive method in science and the requirements of intersubjective agreement in ethics has been developed extensively in Rawls's pre-instrumentalist paper, "Outline of a Decision Procedure for Ethics," *Philosophical Review LXVI* (1957), 177-197. The extent to which he has retained in his later writings a commitment to the value of objectivity developed there is explored in Chapter X, below.

prescribes instrumentally promote an agent's ends *whatever they are* implies that they promote not just someone's final ends, but anyone's. Thus as in the case of action, the possibility of supplying objective evidential support for one's favored moral theory is retained, by exploiting its instrumental connection with theoretical rationality.

Instrumentalism as I have described it characterizes in a very general way a large variety of justificatory strategies that differ considerably in their details from case to case. For example, I have claimed that Instrumentalism attempts to justify a moral theory as the best means to a wide range of unspecified ends. But different moral philosophers impose different structural constraints on that range, and thus decide differently how wide that range can be. For Hobbes, an agent's relevant range of final ends to which the Laws of Nature are claimed to be instrumental are circumscribed by the existence of other agents who are more or less equally strong, intelligent, and self-interested.[19] For Sidgwick, the final ends to which commonsense moral precepts are claimed to be in fact instrumental are those definitive of utility, understood as an internal, independent state of pleasurable consciousness that all agents are presumed ultimately to desire.[20] For Brandt, as we see in Chapter XI, the final ends that an Ideal Code-Utilitarian society is argued to promote are those that would survive cognitive psychotherapy, understood as a process by which one's desires are maximally corrected by vividly represented facts and logic.[21]

Similarly, different moral philosophers impose different motivational constraints on the agent assumed to choose the moral theory. For Gauthier, the choosing agent must be transparent in the sense that others are able to detect any insincerity in her commitment to conform to the precepts of morality.[22] For Harsanyi, the choosing agent must assign an equal probability to occupying any social position under the set of social arrangements that results from implementing the chosen theory.[23] For Rawls, in a later revision of his views in *A Theory of Justice*, the agent is presumed to be overridingly motivated by the desire to realize and exercise her capacity for an effective sense of justice;[24] and so on.

---

[19]Thomas Hobbes, *Leviathan*, Ed. Michael Oakeshott (New York: Collier Books, 1977), Chapter 13.
[20]Henry Sidgwick, *The Methods of Ethics* (New York: Dover, 1966), Book IV, Chapter III.
[21]*Op. cit.* Note 15.
[22]David Gauthier, *Morals by Agreement* (New York: Oxford University Press, 1985), Chapter VI.
[23]John C. Harsanyi, "Morality and the Theory of Rational Behavior," *Social Research 44* (1977), 623-656.
[24]John Rawls, *The Dewey Lectures 1980*, "Rational and Full Autonomy," *The Journal of Philosophy LXXVII* (1980), Section IV. I mention this version of Rawls's view here in order to mark that Rawls's commitment to Instrumentalism has survived his

These moral philosophers have in common that they conceive neither the circumstances under which moral principles are chosen, nor the ends relative to which they are taken to be instrumentally justified to be absolutely unlimited. They each suppose that some constraints must be imposed both on motives and on ends in order for the right kind of choice to be made. In what follows, we will see that there is good reason for this shared supposition. However, we will also see that, at least relative to Instrumentalism as I have characterized it generally, no such constraints can succeed in providing an objective moral justification for any viable normative theory. If this argument is sound, it has significant implications for any normative view that deploys the Instrumentalist strategy. I consider two such views in the two chapters subsequent to this one.

4.2. Justification

The appeal of the Instrumentalist strategy, even in the very general form stated above, is clear. Prereflectively we may suppose that which action is instrumentally rational depends entirely on the very specific further, final ends a particular agent wants to achieve. This supposition implicitly equates rational justification with correct practical reasoning. From this vantage point, we may find initially mystifying the suggestion that some actions are objectively justified instrumentally regardless of the particular character of one's final ends. But on further reflection, we can appreciate the plausibility of this suggestion.

Take for example, behaving courteously. Sometimes behaving courteously has clear disadvantages. Among many others, it frustrates opportunities to vent one's irritation or to demonstrate one's lively wit at someone else's expense. Nevertheless, it might be claimed that it pays to behave courteously to others no matter what. First, one can vent one's irritation just as well by kicking a pillow; and demonstrate one's lively wit at someone else's expense with whoopee cushions, water-squirting lapel flowers, and the like. Besides, just how much of a buzz is it possible to obtain by gratifying one's impulse to be rude? Moreover, others will be more positively disposed toward one, and so more positively disposed to help one further one's ends if one is courteous than if one is abusive, as long as one's ends do not seem to threaten theirs. Furthermore, one cannot know in advance who will be in a position to help or hinder the achievement of one's ends. Since one loses so little by restraining one's impulse to verbal abuse, it pays over the long term to behave courteously to everyone, whatever other ends one may have. Behaving courteously, then, would seem to be an action that is

---

abandonment of many of the characteristics stipulated of the Original Position in *A Theory of Justice*. I discuss the implications of this in Chapter X, below.

instrumentally rational for a very wide range of ends, and so objectively justified to that extent.

Note that the intuitive appeal of the above reasoning depends on two connected features. First, the ends to which having courteously are instrumental are assumed to be motivated by the desire to achieve them. This is true by definition, relative to the Humean model of motivation. For on this model, the only thing that can motivate action in the service of some end is a desire for that end. On the Humean model, if I am motivated to achieve an end, i.e. if it is really my end, then I have a desire to achieve it.

The second, connected feature of the above justification is the substantive weakness of the resulting constraints. The sole motivational constraint is that one has a desire to promote one's ends. The sole constraint on those ends is that they do not appear to threaten the ends of those to whom one is to behave courteously. These constraints leave open to an impressive extent the substantive nature of the ends that may be promoted by behaving courteously, and so the substantive motivation of any agent who may be persuaded to do so. They give everyone whose reasoning is accurately described by such a justification a reason to behave courteously.

So the above argument counts as a candidate for an objective justification of behaving courteously and not just as a bit of correct practical reasoning contingent on the particular ends an agent happens to have, because the argument in question gives each of us, as audience, a reason for adopting this as a rule of conduct irrespective of the particular antecedent ends each of us happens to have. A reason for your adopting this action as a rule of conduct – a reason that is assumed to approximate objective validity as the number of agents for whom it is a reason increases – is not just that it promotes your ends; this would make it merely your reason. An objective reason for you to adopt it is – as Nagel has shown – that it promotes everyone's ends. Hence its status as a reason, to that extent, does not depend on the particular antecedent ends you happen to have. It is an objective reason precisely to the extent that it is everyone's reason.

Of course the fact that behaving courteously promotes everyone's ends cannot constitute an objective *moral* justification of behaving courteously. For among the ends that behaving courteously promotes may be recognizably immoral ones; as when, for example, I behave courteously because this enables me to accumulate political favors which I then cash in for the purpose of ruining my enemies. If immoral ends of this kind, too, are among those which behaving courteously promotes, and if part of the persuasive appeal of behaving courteously is its all-purpose character, then this argument supports the pursuit of immoral ends. This means that the Instrumentalist strategy cannot yield a moral justification of a moral theory, if it shows the actions or set of social arrangements that the theory prescribes to be instrumental to the

promotion of just *any* ends, including recognizably immoral ends. In this case it may justify the theory, without morally justifying it.

It seems, then, that we cannot generate a specifically moral Instrumentalist justification of an action or set of social arrangements without imposing or presupposing at least some prior moral constraints on the range of ends the choosing agent is assumed to desire to promote – as the moral philosophers mentioned above all seem implicitly to recognize. Next I suggest that to the extent that such constraints are imposed, either the action or set of social arrangements in question cannot be justified, or else the Humean model of instrumental rationality is doing no work in justifying them.

4.3. The Incredible Shrinking Means

Now consider a second example of an action that is instrumental to certain final ends, namely giving one's money away. Consider what an instrumentally rational justification of this action might look like, keeping in mind that such a justification must attempt to persuade, not just some few agents, but everyone, you included, that it is rational to give one's money away. Giving one's money away seems to have certain obvious disadvantages. It may frustrate one's opportunities to indulge expensive tastes, or to satisfy certain desires for which money is a prerequisite, such as buying one's parents a house or securing a high-quality education for one's children. It also leaves one in a position of relative insecurity, for one cannot know in advance what emergencies the future may bring. Unfortunately there seem to be no obvious compensations for these disadvantages.

However, this depends on the kinds of desires one has. If one takes one's expensive tastes very seriously, or are particularly committed to securing the wellbeing of one's family, or to being prepared for future emergencies, then the disadvantages of giving one's money away may seem practically insurmountable. But if one does not happen to care as much about these things as one does about supporting anti-racist initiatives, ending worldwide famine, and fighting cultural imperialism – let us call these beneficent ends – then the disadvantages may be more than adequately outweighed by the range of ends one cares about that giving one's money away enables one to promote. So if one has beneficent final ends, and one agrees that giving one's money away is the best way to promote them, then one has a reason for giving one's money away. If you have such a reason, giving your money away would seem to be instrumentally rational for you.

Of course if you do not happen to have beneficent final ends, then you will not be as impressed by the argument that giving one's money away enables one to realize them. Not only will you fail to be persuaded by this argument. You may not even recognize it as an *argument*. Rather than an argument or attempt at justification, this claim may strike you as little more

than an observation, i.e. a bit of correct practically reasoning contingent on the particular ends some other agent may happen to have.

Certainly this reasoning may be supplemented by further argument to the effect that you *ought* to be the kind of person for whom such beneficent final ends outweigh other kinds. You may or may not find such arguments persuasive. If you do not, you will need to be persuaded that you *ought* to *want* to be this kind of person; and if not by this argument, by an argument that you *ought* to *want* to *want* to be this kind of person; and so on. You will need to be persuaded, at some point in the regress, that you have some terminating, rationally authoritative obligation, however tenuous, that links you in your present state to the promotion of beneficent final ends, in order for you to recognize the promotion of beneficent final ends as a justification for giving one's money away. But even if you do so recognize them, it is hard to see how any of these latter arguments will succeed in justifying to you *your* giving *your* money away, if you do not in fact have beneficent final ends. For they will not demonstrate the instrumental rationality of that action to any end you actually have.

So the success of the Instrumentalist strategy depends on the inclusiveness of the range of ends to which the prescribed action or set of social arrangements is in fact instrumental. In order to insure the unanimity of the choice among actions or sets of social arrangements, all agents must be presumed to share final ends toward which the prescribed action or set of social arrangements is instrumental. If the prescribed action or set of social arrangements requires a significant degree of beneficence or altruism, the agents' final ends must be characterized accordingly; and if not, then not. In either case, only if your ends are among them will it justify that action or set of social arrangements to you. And only to the extent that most people's ends are similarly among them will that justification seem to approximate objective validity.

Thus Instrumentalist moral philosophers have two choices. They may water down their normative prescriptions for action to the point of providing little more than a rationale for the status quo, in order to secure for those prescriptions the largest number of final ends to which they are instrumental; Sidgwick would be a classic example of a philosopher who approaches objectivity at the expense of normative substance. Or they may issue normative prescriptions that may require rather a great deal of personal self-improvement in order for most people to follow, and simply stipulate that they speak of and to only that much smaller range of agents who share such ultimate aspirations to begin with. Rawls' more recent views would exemplify this alternative.

But the smaller the range of ends promoted by the action, the fewer the individuals likely to hold them, and the less the instrumentalist justification will approximate objectivity. Call such an action or set of social arrangements

a *shrinking means*. A shrinking means presents an obstacle to supplying an objective moral justification of an action or set of social arrangements, to the extent that the sympathetic audience it selects is correspondingly esoteric. To this extent, it thwarts the Instrumentalist attempt to demonstrate that reason directs *all* of us to perform certain actions in order to achieve *any* of our given ends, and replaces this with what we might describe as an exclusivist demonstration that reason directs *some* of us – those of us with, say, beneficent ends – to perform certain actions in order to achieve *some* of our given ends. Thus it diminishes the Instrumentalist strategy of objectively justifying certain prescriptions to a mere piece of correct practical reasoning that is contingent on certain ends some of us may happen to have.

Now consider a natural Instrumentalist response to the problem of the shrinking means. The response is, essentially, to retort that we cannot concern ourselves with those who do not share our ends, for they lie outside our moral community. If a person does not care about being beneficent (say), then there is nothing more we can say to persuade him of our favored moral theory. We must, it is claimed, suppose ourselves to be talking to those whose basic values are at least roughly similar to our own.[25] The difficulty is that to the extent that this is true, either the action or set of social arrangements in question has not been justified, or else the Humean model of instrumental rationality is doing no work in justifying them.

The action or set of social arrangements in question has not been justified because our acceptance of it is now contingent on having certain particular antecedent ends. If we do not happen to have, say, beneficent ends, or if our ends gradually become less beneficent, perhaps as we get older and familial responsibilities encroach on us more and more, then the prescribed action will become correspondingly contingent and dispensable. This is an acceptable feature of an agent's practical reasoning about particular ends and how to achieve them: if one stops wanting the end, one no longer has reason to perform the action instrumental to achieving it. But this is less acceptable in reasoning that purports to furnish an objective justification of an action or set of social arrangements. For as we have already seen, what makes a piece of reasoning a candidate for an objective justification is its ability to give us a reason for adopting an action or set of social arrangements *independently* of the particular antecedent ends we happen to have. But restricting the appeal of this reasoning to those who must be presupposed to share our particular antecedent ends violates this criterion. An action or set of social arrangements cannot be both objectively justified and a shrinking means. Restricting the

[25]See, for example, Phillippa Foot, "Morality as a System of Hypothetic Imperatives," *The Philosophical Review LXXXI* (1972), 306-16; Gilbert Harman, "Moral Relativism Defended," *The Philosophical Review LXXXIV* (1975), 3-22. Rawls also seemed to move in this direction. See his *Dewey Lectures* (*ibid.*, 537).

range of final ends to those to which a preferred action or set of social arrangements is instrumental means abdicating the aspiration to the objective validity of the theory that prescribes it.

4.4. The Problem of Moral Justification

The contingence and dispensability of a shrinking means is a liability for an objective justification of it. But for a purportedly objective *moral* justification of it, its contingence and dispensability is a quite fatal liability. For an objective moral justification of an action or set of social arrangements is supposed to persuade us that we ought to observe its implied prescriptions *whatever else we do*. That is, an objectively valid moral theory is supposed to demonstrate its prescriptions to be *absolute* constraints on action, and not mere rules of thumb contingent on the particular antecedent ends some of us happen to have. Indeed, if the theory is to provide absolute constraints, even a shrinking means sufficiently comprehensive to promote everyone's *de facto* antecedent ends will not do the trick. For even here, its promoting everyone's ends supplies me with a reason to act on it that is independent of any of my antecedent end thus promoted only because of the particular antecedent ends it does promote, namely everyone else's. This means that, at best, the Instrumentalist strategy can justify an action or set of social arrangements independently of *any particular* antecedent end it promotes. Instrumentalism cannot justify an action or set of social arrangements as objectively valid independently of *all* antecedent ends, i.e. absolutely. So Instrumentalism can approximate but cannot achieve objective moral justification, because any action or set of social arrangements it attempts to justify must function as a relatively shrinking means, however inflated it may seem.

But an absolute moral justification is needed, so that we can make the kinds of moral judgments a moral theory should enable us to make. A normative moral theory does not direct us to respect others, to behave responsibly, to help the needy, and to be honest in our dealings only when it is convenient and not otherwise. The whole point of a normative theory is to guide behavior correctly in those cases where self-interest obscures the morally right thing to do. Similarly, a normative moral theory is supposed to enable us to make negative moral judgments about actions or sets of social arrangements that violate the prescriptions implied by our theory. But in order to be *moral* judgments, these judgments cannot find the action or set of social arrangements defective simply because it does not best promote, say, the beneficent ends we are presumed to share. Such a judgment would not be a moral judgment but rather a judgment of practical irrationality. In order to be a moral judgment, it must evaluate the action or set of social arrangements as right or wrong *independently* of our particular antecedent ends. It must be able to make judgments about the actions of agents who do not share our beneficent ends and values – for example, that an agent does right to

contribute to charity even though she wants only thereby to enhance her social standing; or that an agent does wrong to keep promises in order to accumulate political favors which he may later cash in for the purpose of ruining his enemies. We cannot simply throw up our hands in such cases and decline to judge the behavior of such agents on the grounds that they do not share our ends and therefore are not members of our moral community – as though a moral community were nothing but a private club or chat room (who might comprise a moral community is discussed at greater length in Volume II, Chapter X). A moral theory must be able to judge when such a person's actions are wrong because the ends that define and identify those actions are wrong – objectively, morally wrong. If a moral theory is not objectively valid in this sense, it is unclear why anyone would have reason to hold it.

Now put this problem aside for the time being. Assume we can go on thinking of a shrinking means as objectively justified, despite its esoteric appeal, to those who have, say, beneficent ends. I shall signal this assumption henceforth by putting "justify" in scare-quotes when using it to refer to a shrinking means. Here I am assuming what is false, namely, that our moral community consists only in those who share our central ends and values. In this case the Humean model of instrumental rationality is in any case doing no justificatory work. For what "justifies" my giving my money away is not the fact that it is the most efficient means to my beneficent ends; as we have already seen, this consideration does not differentiate a piece of correct practical reasoning from an objective moral justification. Nor is what "justifies" my giving my money away the fact that everyone in my moral community shares the beneficent end to which this action is instrumental; as we have also seen, this contingent intersubjective unanimity is not what guarantees the objective validity of giving my money away.

Rather, what "justifies" my giving my money away is the fact that, as we have already seen in Chapter V.3, this action itself can be regarded as *constitutive* of beneficence. The point has more general application, but let us confine it to the case of morality. It holds of any action claimed to be a means to a set of ends characterized in morally specific terms. A set of ends is *characterized in moral specific terms* if normative terms (such as "good," "fair," "beneficent," "evil," "selfish," "unjust") are among the predicates we ascribe to each member of the set. If we ascribe moral predicates to the ends we aim to achieve by acting, then since actions are defined and identified by their ends, those selfsame predicates can be applied equally to the actions we take to achieve them. Let us describe such predicates as having a *retrospective* application. So, for example, if my ends are good, then the actions I take to achieve them may be characterized similarly. Of course they may be other things as well, such as stupid, ill-considered, naive, and so forth. If my ends are beneficent, then giving my money away can be described as a beneficent

act. If my end of acquiring as much personal power as possible in order to ruin my enemies is vengeful, then I can be regarded as acting vengefully in accumulating political favors in order to achieve it. These three examples show that moral predicates ascribed to an end can be applied equally to the action taken to achieve it, regardless of how vaguely or specifically either is characterized. It is the value of our ends, not their popularity or the efficiency with which they are pursued, that confer value on the actions we take to achieve them.

This may not seem obvious. It may be objected that, for example, if I have the virtuous end of improving social relations among my colleagues at work, and a necessary means to that end is that I dress warmly before going to work in the morning, it does not follow that my action of pulling a second pair of woolly socks over my feet is virtuous. But my claim is neither that any such action must be so characterized, nor that it cannot be characterized alternatively. My claim is simply that it can be so characterized, in so far as it is understood as promoting the good end in question. However, this conclusion does not extend to just any terms in which an agent's ends are characterized. For example, it does not follow from the fact that my ends are varied that the actions I perform in their service are varied as well. Nor should it be thought that the morally specific terms that characterize an action necessarily have *prospective* application to its end. From the fact that behaving courteously is morally virtuous it does not follow that all the final ends it promotes can be characterized as morally virtuous as well. But this asymmetry is to be expected. For part of what we want to say is that some actions are susceptible of moral evaluation *independently* of the further ends they promote. The problem with an Instrumentalist strategy that uses a shrinking means is that it does not allow us to say this.

If a shrinking means can always be regarded as constitutive of the moral end it promotes, then as we have just seen, its status as an efficient means to that end cannot be what "justifies" it. It is rather the value conferred on it by that moral end itself that does the justificatory work. Indeed, the whole point of imposing moral constraints on the range of ends an agent is assumed to desire to promote via the action or set of social arrangements in question is to *subordinate* efficiency considerations to moral ones. This implies that moral considerations are overriding in evaluating the suitability of means to our moral ends.

So we who share that moral end are not persuaded to adopt a shrinking means because it *efficiently* achieves that end. Any action that could be characterized similarly in terms of it would have the same persuasive force. For example, even if distributing fliers promoted our beneficent ends less efficiently than giving our money away, that they did so would "justify" distributing fliers just as well. Certainly we might want to invoke considerations of efficiency in choosing between the alternatives of giving our

money away and distributing fliers, if we could not do both. But in this case the primary efficiency considerations would ordinarily concern which alternative was less costly *to us*, given our other ends. They would not, unless we were martyrs or efficiency fanatics, concern which was less costly *tout court*. It is not even clear what this would mean.

But if the primary efficiency consideration concerns which action is least costly to us rather than which is least costly *tout court*, then according to the Humean model of motivation, my choice of giving my money away rather than distributing fliers (or vice versa) *makes* that choice the most efficient action for me to take, for any action I choose. So it would be the fact that the action achieved our moral ends, rather than that it did so efficiently, that "justified" that action to us. But in this case, the notion of efficiency that is centrally definitive of the Humean model of rationality is doing no justificatory work. It is rather the values we hold in common that persuade us to adopt the means for realizing them.

I have discussed two examples of such means: behaving courteously and giving one's money away. We have seen that these two differ in systematic ways. Behaving courteously is instrumental to a wide range of ends. For that very reason, I have suggested, behaving courteously can be objectively justified to a degree, but to that degree cannot be morally justified. By contrast, giving one's money away is instrumental to a more limited range of ends. For that very reason, I have claimed, it cannot be objectively justified to any degree, but to that degree can be morally "justified." There is one further notable difference between behaving courteously and giving one's money away. Behaving courteously is easy. Giving one's money away is hard. It is not surprising that we can be more easily persuaded to do things that are easy than things that are hard, nor that the Instrumentalist strategy is particularly well suited to thus persuade us. This is a consequence of the background Humean conception of the self, according to which we are motivated to do things that efficiently promote ends and values we are already assumed to have. We are not so easily motivated to do things that require us to adopt new ends, and even less so if they require us to modify our values or priorities. Actions we recognize as morally virtuous but hard to motivate ourselves to do are actions for ends we have not seriously adopted, ends that express values to which we may give lip service at best – ends and values that may well lie beyond the *actual* moral community of which we are in reality members.

It would be very regrettable if we could find no moral theory persuasive that enjoined us to do things that are hard, things that required us to modify or sacrifice our ends, because in that case we could find no reason to sacrifice where we are able for the sake of the common good. But if it is in any case, as I have suggested, the values we hold in common that persuade us to adopt the means for realizing them, rather than considerations of instrumental

rationality, then our willingness to sacrifice where we are able for the sake of the common good will depend on the values we hold, and on the conditions under which we can be rationally persuaded to modify them. And then it becomes crucial to ascertain whether those values themselves are rationally justified. As we will see in the following two chapters, to answer this question we will need to press beyond the limitations of Instrumentalism.

Chapter X. Rawls's Instrumentalism

I have just argued in Chapter IX.4 that the Instrumentalist strategy of moral justification is self-defeating. John Rawls and Richard Brandt have each utilized the Instrumentalist strategy in defense of their respective moral theories, by arguing that their theories would be chosen by a fully instrumentally rational agent concerned to further her own ends, whatever these might be, given the available information, whether limited (Rawls) or full (Brandt). However, the success of this strategy requires each to make further, controversial assumptions about the chooser's ends and motivations in the choice situation, in order to derive the favored moral theory. I address Rawls's theory in this chapter and Brandt's in the next.

Common lines of early criticisms of Rawls's *Theory of Justice*[1] made by, among others, Schwartz, Nagel, Gauthier, and Miller were alike in presupposing the *continuity thesis*, i.e. that the parties in the original position are psychologically continuous with members of the well-ordered society. To my knowledge, no later commentators on Rawls's writings have disputed this assumption. There is evidence in *A Theory of Justice, The Dewey Lectures*[2], and other writings by Rawls both to confirm and to disconfirm this thesis. His late *Political Liberalism*[3] appears to dispense with it – without, however, addressing the issue directly. If this thesis is true, then either the original position cannot generate any principles of justice at all; or else Rawls's special motivational assumptions about the parties in the original position, i.e. that they are overridingly concerned to develop and express their moral personalities, and secondarily, to advance their conceptions of the good, tautologically imply that they will choose the principles of justice, in order to distribute primary goods. In this case, it is vacuously true that an agent will choose what he has special motivation to choose, other things equal. But in order to justify this choice for us, we, too, must have that special motivation. But if we do, then the argument does not succeed in justifying this choice as instrumentally efficient *whatever* our ends, i.e. objectively. If we do not, then it does not succeed in justifying this choice for us at all. In either case, the Instrumentalist strategy fails. However, if the continuity thesis is false, then Rawls's early ambition to conceive moral justification on analogy with scientific justification must be reevaluated. Greater attention then must be focused on Rawls's conception of wide reflective equilibrium as a justificatory device, and Rawls is correct in maintaining the irrelevance of the question of personal identity to the construction of his moral theory.

---

[1]John Rawls, *A Theory of Justice* (Cambridge, Mass.: Harvard University, 1971). Henceforth references to this work are parenthecized in the text and denoted by "TJ".

[2] John Rawls, *The Dewey Lectures 1980*, *The Journal of Philosophy LXXVII* (1980)

[3] John Rawls, *Political Liberalism* (New York: Columbia University Press, 1996)

Section 1 grounds the type of justification I want to discuss in Rawls's early and very ambitious "Outline of a Decision Procedure for Ethics,"[4] in which he explicitly sets his sights on a conception of moral justification analogous to scientific explanation. Section 2 offers a general description of the basic features that identify traditional Social Contract Theory. Section 3 situates Rawls's central metaethical and normative views within that context, and Section 4 examines and refutes some of Habermas' criticisms of Rawls's normative and metaethical theories. Section 5 reevaluates what Rawls's metaethical approach to moral justification can achieve relative to the analogy of scientific justification on which it is based, with particular reference to Rawls's threefold distinction among perfect, imperfect and pure procedural justice. Section 6 traces the evidence in Rawls's work for the continuity thesis, and its use by some early critics of *A Theory of Justice*. Section 7 analyzes the Humean and Instrumentalist underpinnings of Rawls's views, and the quite serious contradiction in Rawls's conception of the original position that these assumptions generate. Section 8 considers the resources within Rawls's view for resolving this contradiction and thereby answering his early critics. Section 9 extends the metaphysical implications of this alternative interpretation of the original position to a resolution of the question of personal identity, and also reconsiders the potential of wide reflective equilibrium to provide an alternative to the Instrumentalist strategy of justification. Section 10 spells out the alternative conception of justification as analogous to scientific procedure that the concept of wide reflective equilibrium entails, and concludes with some observations about how Rawls might finally deploy this concept to satisfy the stringent requirements of moral justification which he set for himself in 1951.

1. The Analogy with Science

In 1951 John Rawls expressed these convictions about the fundamental issues in metaethics:

> [T]he objectivity or the subjectivity of moral knowledge turns, not on the question whether ideal value entities exist or whether moral judgments are caused by emotions or whether there is a variety of moral codes the world over, but simply on the question: does there exist a reasonable method for validating and invalidating given or proposed moral rules and those decisions made on the basis of them? For to say of scientific knowledge that it is objective is to say that the propositions expressed therein may be evidenced to be true by a reasonable and reliable method, that is, by the rules and procedures of what we may call "inductive logic";

[4] John Rawls, "Outline of a Decision Procedure for Ethics," *Philosophical Review 66* (1951), 177-197; reprinted in *Ethics*, Ed. Judith J. Thomson and Gerald Dworkin (New York: Harper and Row, 1968), 48-70.

> and, similarly, to establish the objectivity of moral rules, and the decisions based upon them, we must exhibit the decision procedure, which can be shown to be both reasonable and reliable, at least in some cases, for deciding between moral rules and lines of conduct consequent to them.[5]

In this passage Rawls expresses impatience with the traditional treatment of metaethical issues as a branch of speculative metaphysics. He reconfigures the issue of moral objectivity and reorients the practice of metaethics from linguistic analysis to decision theory. By turning attention to the correct procedure for making substantive moral decisions – about what action is right as well as about what kind of society would be good for human beings, Rawls thereby revitalized the practice of normative moral philosophy, and of casuistry more concretely, after a century of relative neglect.

At the time Rawls wrote "Outline of a Decision Procedure for Ethics," moral philosophers did not often address normative moral questions. They were more centrally concerned with the metaethical status of such questions themselves, and were accustomed to couching their concerns about the objectivity of moral judgments in the following terms: Do terms such as "good" and "right" refer to anything? And if so, to what do they refer? Abstract entities? Emotional states of the speaker? Can a moral judgment be objectively true independently of the local moral code in which it figures? If so, in what does this truth consist? If moral terms do not refer, on what basis do we accept the judgments in which they figure as objectively valid? Or are all moral terms and judgments valid only relative to a particular speaker, community, or culture? This last conclusion would entail a correspondingly relativized and reduced role for the normative moral philosopher, and so a contraction in the scope of philosophical ambition of the kind already discussed: If some such form of relativism were true, we would have no moral justification for intervening in any of the practices or behavior of other individuals or groups whose actions we found objectionable. Fighting for civil rights, protesting international human rights violations, and helping the needy would find no more solid legitimation than imperialism, hegemony, or meddling.

In response to this possibility, Rawls replaced Moore's question, Do moral terms refer? with a different one: Can moral judgments be the outcome of a rational and reliable procedure? He argues in the above passage that the conception of moral objectivity on the basis of which these issues traditionally have been framed is itself wrongheaded. The project, he argues, should not consist in a search for abstract metaphysical entities, corresponding to moral terms and judgments, which we can metaphorically pinch, kick and pummel to reassure ourselves of their objective reality as we physically do bodies,

---

[5]*ibid*.

tables and pillows. Nor should it consist in a cross-cultural statistical study of the variety of moral codes that govern various human societies. Instead we need to learn from the procedures of decision-making and verification used in the natural sciences, and from the conception of objectivity defined by those procedures.

We have seen in Chapter IX.4 that in the natural sciences, a judgment is taken provisionally to be objectively valid if it is the outcome of the procedures of inductive logic: observation, adequate gathering of data, inductive hypothesis formulation, deduction of predicted outcomes, testing of those predictions under controlled conditions, and intersubjective replication of predicted experimental results. Objective scientific truth ideally is conceived as a function of rational investigative procedure carried through without error. Similarly, Rawls suggests, objective moral truth, ideally, is a function of a rational decision procedure carried through without error.

In the remainder of this early paper, Rawls proposes such a procedure, and refines it further in the series of publications that succeeded it.[6] But it is not until *A Theory of Justice* that he is prepared unapologetically to defend the thesis that moral philosophy is to be considered "part of the theory of rational choice (TJ 16, 47, 172);" and to assert that "[t]he argument aims eventually to be strictly deductive (TJ 121)." In yet later writings, Rawls had occasion to revise and qualify this stance.[7] In *Political Liberalism*, he decisively disowns it (PL 53, fn. 7). His considered qualification of his earlier enthusiasm about the extent to which moral philosophy could aspire to objective universality is a tribute both to the seriousness with which he took his critics' objections, and to his commitment to the value of Socratic metaethics more generally.

I believe Rawls took his critics' objections a bit too seriously, and did not need to retrench quite as much as he did. Nevertheless, there are serious problems with the Instrumentalist strategy of metaethical justification to which Rawls, like all devotees of the Humean conception of the self, is committed; and in this chapter I examine the particular ways in which this strategy leads his metaethical project astray. My analysis has no implications for the truth or falsity of Rawls's substantive, normative theory of justice. Nor does it imply that this normative theory cannot be metaethically justified. Indeed, I argue, finally, that Rawls's concept of wide reflective equilibrium

---

[6]See his "Justice as Fairness," *The Philosophical Review 57* (1958); "The Sense of Justice," *The Philosophical Review 62* (1963); Constitutional Liberty and the Concept of Justice," *Nomos VI: Justice*, Ed. C. J. Friedrich and John Chapman (New York: Atherton Press, 1963); "Distributive Justice," in *Philosophy, Politics and Society*, Third Series, Ed. Peter Laslett and W.G. Runciman (Oxford: Basil Blackwell, 1967); "Distributive Justice: Some Addenda," *Natural Law Forum 13* (1968); and "The Justification of Civil Disobedience," in *Civil Disobedience*, Ed. H. A. Bedau (New York: Pegasus, 1969).

[7]See, for example, his "Justice as Fairness: Political not Metaphysical," *Philosophy and Public Affairs 14*, 3 (1985), 223-251.

constitutes an alternative rational procedure that is not dependent on the Instrumentalist strategy he in *A Theory of Justice* deploys; and that this procedure can provide the metaethical defense of the theory as objective moral truth that the Instrumentalist strategy does not.[8]

2. Traditional Social Contract Theory

2.1. The Normative Theory

In *A Theory of Justice*, Rawls positions himself as a Social Contract Theorist in the tradition of Hobbes, Locke, Rousseau, Hume, and Kant (some would add Hegel). This is a lengthy list comprised of widely divergent philosophical sensibilities. But almost all share certain fundamental normative beliefs in common constitutive of the doctrine of *liberal democracy*. Most generally, almost all believe we should have a society that maximizes individual freedom to pursue personal goals and interests, given the existence of other people. Almost all assume, additionally, that the existence of other people, all equally engaged in pursuing personal goals, requires constraints imposed by a governing body that regulate and coordinate orderly interactions among them (here Hegel would be the clearest exception). From these two assumptions almost all conclude that society ideally should be structured so as to protect individual rights, freedom and autonomy as fully as possible, i.e. so as to maximize political equality, consistently with these constraints (Hobbes' view would necessitate qualification of this claim). Political equality, then, for the Social Contract Theorist, means equal protection under laws designed to safeguard these freedoms. It means that no interference by the governing body in the rights of individual citizens to conduct their affairs as they choose that is not justified by the ideal of political equality itself is defensible in a court of law; and that any individual citizen can call upon the legal system to protect her against any such infringement.

The ideal of liberal democracy is therefore characterized by an in-principle refusal to specify the content of personal goals or values deemed acceptable or worthy of pursuit by individual citizens. Individuals are assumed to be able to decide these matters for themselves, and to have the right to do so. Attempts by the state to define, impose, or circumscribe the range of such goals, beyond the bare minimum required for social cooperation and stability, are taken to be unwarranted interference in the exercise of

[8] Rawls accords greater significance to wide reflective equilibrium along these lines in explicit response to Habermas' critique of his views. See Jürgen Habermas, "Reconciliation through the Public Use of Reason: Remarks on John Rawls's Political Liberalism," and Rawls's "Reply to Habermas," both in *The Journal of Philosophy XCII*, 3 (March 1995). Rawls's reply is reprinted in *Political Liberalism* and I use that pagination. See particularly PL 384-385.

individual liberty. Whereas the state can compel citizens to observe restrictions on liberty required in order to maximize liberty overall, it cannot compel them to have, profess, or act upon any other such values.

In this sense the normative value content of traditional Social Contract Theory is deliberately minimal, for the value of liberty itself does not enter into the hierarchical prioritizing of multiple values among which an individual may make rational trade-offs (for example, assigning a higher priority to satisfying work than to a relaxing vacation). Instead it functions as a *side-constraint* on the set of values any such individual may have: Whatever their content, all must be consistent with respect for individual liberty. No other constraints on the content of individual goals or values are permitted. So, to invoke Rawls's famous example, someone who solves complex mathematical equations for a fee so that he can maximize the time he spends counting blades of grass (TJ 432) cannot be prevented from pursuing this goal on grounds of futility, sloth, or lack of productivity. If counting blades of grass is the activity he values most and he does not waste any scarce social resources in its pursuit, he is free to do so.

Also implicit in this in-principle refusal to evaluate the content of individual goals and values is a subjectivism about what constitutes acceptable or worthwhile ones. Answers to this question are left to the individuals in question – not only because individuals alone have the right and the liberty to answer them for themselves; but also because no individual is granted the authority to prescribe value for any other. There is no equivalent in social contract theory to the explicitly authoritarian roles of Pope, Ayatollah, or dictator, whose actions and/or pronouncements function as symbolic embodiments of personal ethical, political, or spiritual values held in common by all. Instead the rulings of the governing body are conceived as at most enacting a "General Will" to protect liberty itself.

Traditionally, the Social Contract-Theoretic ideal of political equality is a juridical ideal. It does centrally affirm the right to private property. But it is famously silent on the question of how economic resources are to be distributed among individual citizens. This follows from its hands-off attitude toward the content and worth of individual goals and interests. Since it makes no assumptions, within the constraints imposed by the requirements of social cooperation and stability, about what goals and interests are worth pursuing, it similarly makes no assumptions about the quantity or quality of resources instrumentally necessary to pursue them. Therefore there can be no justification inherent in the traditional conception of the social contract itself for stipulating any basic social minimum to which all citizens are entitled, since any such minimum would arbitrarily presuppose a particular range of ends whose pursuit would require it. Since no such presupposition can be justified in traditional Social Contract-Theoretic terms, no such social minimum can be, either. The traditional contract-theoretic view implies that

those who find shelter under a bridge or in a subway tunnel, and sustenance in a soup kitchen or garbage bin implicitly choose to do so; and reject the concept of a social minimum as an unwelcome burden, imposing unwelcome obligations, in a lifestyle choice that maximizes freedom of mobility at the expense of security and safety.

2.2. The Metaethical Justification

The Social Contract-Theoretic ideal of liberal democracy is a normative moral theory of the good society. Traditionally its metaethical justification has taken a certain form most evident in the views of Hobbes, Locke, Hume, and Rousseau. In Kant and Hegel this strategy is virtually absent. It consists in arguing, basically, that this ideal of liberal democracy can be conceived as the outcome of an agreement among individuals in a pre-societal "state of nature" to abide by its laws, rules, and norms. Each individual is motivated to obey these norms by her desire to maximize the achievement of her individual goals and interests as efficiently as possible, given the desire of other individuals to do the same. Obedience to the rule of law is defended as the necessary price of protection by the state against the potential for confusion, conflict, disequilibrium, and the consequent thwarting of individual interests and desires that would obtain if no such rule of law existed. The reasoning is that, given our individual desires to achieve our goals, whatever they are, the system of laws and norms that constitute the ideal of liberal democracy is one we each would have consented to, or that was consented to, under pre-societal conditions of disequilibrium.

The strategy of metaethical justification of the social contract is, then, *backward-looking* rather than *forward-looking*. It conceives the social contract, and the ideals of liberal democracy it expresses, as the result of prior conditions imposed on equal and rationally self-interested choosers who begin on a level playing field. That the social contract is the result of these prior conditions is what legitimates it. Traditional Social Contract Theory does not argue that our obeying these rules will in fact have the best consequences for our attempts to satisfy our desires. It merely claims that rationally self-interested individuals would – or did – justifiably think it would, under the condition that they themselves had no rules at all. All traditional Social Contract Theory claims is that if there were no rules governing legitimate social exchange, these are the ones equal, instrumentally rational, self-interested individuals would agree to. That such individuals in a pre-societal state might justifiably reason that obeying these rules will maximize their chances of satisfying their desires does not, of course, imply that our actually obeying them will in fact have this effect.

A fortiori, traditional Social Contract Theory does not argue that the actual individual or collective consequences of implementing the social contract are necessarily superior to those of any other social scheme. Indeed, it could not

argue this. Actual individual consequences must vary according to the individual goals and interests one happens to have and the economic circumstances in which one finds oneself. So, for example, a homeless person living under a bridge who aspires to the U. S. Presidency will probably fare less well relative to his aspirations than the well-trained daughter of a wealthy financier who aspires to continue the family business will relative to hers. And since traditional Social Contract Theory stipulates that there are in principle no collective goals it aims to maximize over and above the freedom to pursue individual ones, there are no actual collective consequences that might demonstrate any such superiority. To reproaches for permitting radical inequalities in wealth and social well-being among citizens, divisive and destabilizing social conflict among groups of citizens, political corruption, neglect of the neediest, or exploitation of the powerless, the traditional Social Contract Theorist responds with the "Yes, but at least ..." argument intended to remind us of the side-constraining value of individual liberty itself. The claim is that any such social ills are more palatable under conditions of political liberty – indeed, more palatable than their amelioration under conditions of political repression would be.

Similarly, the traditional strategy of metaethical justification of the social contract is *Instrumentalist* rather than deontological (again Kant and Hegel are the exceptions here). It does not argue that the implementation of the social contract is a good in itself, or that the rules and norms observance of which constitute the ideal of liberal democracy have intrinsic value as, for example, an expression of human community, or as the culmination of human rationality. It claims, rather, that the social contract is justified by the conditions that generated it, as an instrumental means to the realization of certain further conditions. Suppose it does not, in fact, succeed in maximizing the ability of individual citizens to satisfy their desires. Then either the circumstances under which it was chosen were so different from the ones that obtain now that the instrumental considerations that justified it to the rational individuals who chose it then are similarly different from what justifies it to the agents who abide by it now; or else the individuals who originally agreed to it may have been mistaken in their reasoning. Perhaps they may have lacked relevant information, or made false inferences. In either case, its Instrumentalist justification is accordingly called into question. If the social contract is not justified instrumentally, it is not justified at all.

However, this Instrumentalist justification can be called into question without being refuted, since, as we have just seen, the metaethical justification of social contract theory is Instrumentalist without being "consequentialist" ("consequentialism" is just one variety of Instrumentalism). Unwanted and unanticipated actual consequences of implementing the social contract in practice can call its instrumental rationality into question without thereby requiring the conclusion that it would be rational to abandon it. For it may be

true, on the one hand, that it does not, in practice, maximize the satisfaction of individual desire for all or even most citizens; and, on the other, that it was, after all, the most rational choice under the circumstances in which it was chosen. If those circumstances themselves have special significance to the agents who abide by it now, then the instrumental reasoning that generated it then may survive retrospective scrutiny. The social contract would still be instrumentally justified even though it did not have the best consequences in practice, because the particular circumstances and instrumental reasoning that generated it then would themselves have intrinsic value for us now.

The metaethics of Social Contract Theory as traditionally conceived, then, has certain identifying characteristics. First, it includes a characteristically Humean conception of the self. It conceives of human agents as motivated by the desire to pursue and achieve self-interested goals and values, in whatever these may consist. These goals are conceived as prior to and independent of established political and legal institutions, and as instrumentally justifying those institutions. Human agents are thus conceived as expressing themselves through their capacity for instrumental, means-end reasoning, i.e. their ability to seek out and exploit the most efficient means available for satisfying their desires. In these ways, human agents are also conceived as more or less equal in physical and mental abilities, such that each has a certain minimum degree of strength and calculating ability, such that deficiencies of one are usually compensated by excesses of the other.

Second, the metaethics of Social Contract Theory traditionally includes a preconditioning set of circumstances – the "state of nature" – that, by generating the social contract, thereby justifies it. By definition, these preconditioning circumstances are such that the agreed-upon social arrangements that constitute implementation of the contract are not yet in place. Since individuals are nevertheless motivated to pursue the satisfaction of their desires, these preconditioning circumstances are characterized by confusion, conflict, and mistrust, and this thwarts or obstructs each individual's ability to satisfy his desires.

Third, there is, of course, the social contract itself. Using instrumental reasoning, each individual in the preconditioning circumstances concludes that mutual cooperation with other similarly self-interested agents is the best means for her to pursue her own self-interested goals. Each therefore agrees on terms of mutual cooperation – laws, social norms, and moral rules coordinating expectations and behavior, and providing for sanctions and means of enforcing them – which each also thereby agrees to obey. In order to implement this agreement, individuals voluntarily abdicate a certain amount of personal freedom and power available in the preconditioning circumstances to a mutually agreed-upon governing authority. The function of this governing authority is to regulate and insure obedience to these norms

by all individual parties to the agreement, and so to protect a certain degree of social stability, order, and individual freedom for all.

Thus each individual trades unlimited – but insecure and unstable – freedom and power in the state of nature for a more limited but secure and stable freedom and power under the social contract. Limited but secure and stable freedom and power under the social contract are instrumental goods deemed more efficient means for pursuing one's ends than the alternatives available in the state of nature. Social Contract Theory ranks unlimited power for everyone in the state of nature lower in priority than the benefits of instrumental rationality, and so gives a higher priority to the limitations, constraints and regulations of power that instrumental rationality requires. It answers Nietzsche's devaluation of the character dispositions of rationality with the argument that power is exercised most effectively only with their help.

3. A Theory of Justice

3.1. Rawls's Metaethics

3.1.1. The Original Position

Rawls's metaethical justification of the social contract bears comparison with the traditional one only in certain respects. It is roughly similar in its conception of human nature and agency in the preconditioning circumstances. Here, too, human beings are more or less equal, motivated by desire, instrumentally rational, and primarily concerned to advance their self-interest, that is, they instantiate the Humean conception of the self. However, Rawls introduces a broader conception of self-interest as a *conception of the good*. By "good," Rawls means basically "what it is rational for someone with a rational life plan to desire" (TJ 399, 405). A rational life plan is one that is consistent with the principles of rational choice (TJ 410-416) and chosen with full deliberative rationality (TJ 408). He distinguishes between interests *in* the self – i.e. egoistic desires as traditionally understood, and interests *of* the self, of which egoistic desires are only a subset (TJ 127). Other interests of the self might include other-directed interests such as particular moral commitments, religious convictions, or altruistic social concerns that define an individual self without being directed at the individual self. Thus individual conceptions of the good might still conflict even though they are not primarily egoistic.

By stipulating that the parties in the preconditioning circumstances are also *mutually disinterested*, i.e. that they take no interest in one another's interests (TJ 13, 127-130, *passim*), Rawls insures that even though each may hold a conception of the good guided by other-directed concerns, the others to whom these concerns are directed do not include the other parties in the original position. This enables Rawls to conceive the parties as having

conflicting conceptions of the good without requiring them to be self-interested in the narrow, egoistic sense. It also means that the parties will have no motivation to form special interest groups or power blocs among themselves that might advocate jointly held, other-directed conceptions of the good outweighing other parties' individually held conceptions.

Rawls's version of the preconditioning circumstances in which such individuals find themselves is equally distinctive. The *original position*, as Rawls calls it, is entirely hypothetical. It legitimates a set of social arrangements as just if it is governed by principles of justice we would have chosen if we ourselves had occupied in the original position. There is no suggestion that the original position might have existed historically, or even that it metaphorically describes our actual social relations. So, in particular, Rawls does not claim that actual human agents, embedded in social and political relationships, adhere to moral and social conventions for reasons of self-interest even in Rawls's broader sense of that term. The implication, on the face of it, is that actual human agents may be motivated by conscience, habit, peer pressure, or individual biochemistry, or any of the other myriad motivations that move us. The original position is rather an ideal choice situation counterfactually conceived such that, if it were to obtain, would generate principles of justice. Thus it provides an Instrumentalist justification of those principles even in the event that adhering to them now, under actual circumstances, does not in fact promote our actual self-interest more efficiently than any other.

Rawls's argument in outline is that if the parties choosing in the original position were free, equal, perfectly rational, morally impartial, and motivated to settle conflicts among their respective conceptions of the good, they would choose a *well-ordered society*, i.e. one structured by Rawls's *two principles of justice*. That they would so choose then provides a criterion against which to evaluate the justice of our actual situation (TJ 13, 16, 17). This is what he means by declaring moral philosophy to be part of the theory of rational choice. In *A Theory of Justice*, Rawls's conception of the original position and the parties in it is a decision-theoretic conception that, like decision-theoretic proofs, imposes certain formal and value-neutral conditions on an ideally rational chooser, such that the outcome chosen can be formally derived from the set of those conditions taken as premises. With this reconceptualization of the traditional Social Contract Theorist's state of nature, Rawls brings the arguments for particular conceptions of social justice to a degree of formal rigor that had never before been attempted within the Social Contract-Theoretic tradition, and that has enabled it to compete with the welfare economics of Utilitarianism. This is only one reason why I vehemently oppose Rawls's later disavowal of this key tenet of his theory.

What motivates the parties to choose principles of justice is what Rawls calls the *circumstances of justice* (TJ 126-130). These are of two kinds. The

subjective circumstances of justice consist in the mutually conflicting goals and values of the parties. This, as we have seen, is built into the original position by their mutual disinterest and differing conceptions of the good. The *objective* circumstances of justice consist in a moderate scarcity of resources, such that there are not enough for each individual to have as much as he wants. Only under conditions of moderate scarcity can questions of justice arise: An overabundance of resources presents no need for raising questions of fair distribution (as, for example, fresh air and clean water did not before the industrial revolution); whereas an extreme scarcity of resources – i.e. one in which proportional amounts to everyone is insufficient for each and withholding distribution to anyone is unjustifiable – makes questions of fair distribution impossible to answer. Together, the subjective and objective circumstances of justice – the necessity of fairly distributing moderately scarce resources among individuals with divergent and conflicting plans for their use – constitute the conditions under which such individuals are moved to find principles of such distribution on which they all can agree.

3.1.2. The Parties' Psychology

The parties in Rawls's original position are defined by certain further special psychological characteristics in addition to being mutually disinterested. Rawls's detailed treatment of these and other motivational assumptions in *A Theory of Justice* rekindled attention to a philosophical tradition of moral psychology that had begun with Aristotle but had lain virtually dormant in the analytic tradition after Kant. Rawls stipulates that the parties are not moved by envy, i.e. by rancor and spite (TJ 538) such that they so strongly desire that another have fewer resources that they are willing to accept fewer resources themselves in order to achieve this (TJ 143). This means that individuals do not compete with one another or measure their own degree of well-being comparatively. An envious person, on Rawls's view, is one who engages in the practice of cutting off her nose to spite her face. As we saw in Chapter II.2.4, she regards another's gain as by definition her own loss, which she will take steps to prevent – by choosing to incur a more serious yet more bearable loss for herself. For such a person, no loss is worse than the loss of self-regard she experiences as caused by another's success; and any other alternative loss that will prevent this is preferable. For example, an envious colleague might attempt to thwart one's professional success in order to avoid feelings of personal and professional inferiority, by placing administrative or procedural obstacles in the path of one's research – even though the completion of one's research would yield needed practical and social benefits and precedents for the status and reputation of one's department overall and so for that colleague himself.

Thus an envious person is one who is willing to accept quite substantial social or material losses in order to avoid the feeling – to which she is

especially susceptible – that she is in some respect inferior to another. To say that the parties in the original position are not moved by envy is thus to say that each individual chooses principles of justice in such a way as simply to maximize her distribution of resources relative to the total amount available. Each is guided by a sense of the sufficiency and intrinsic value of their own plan of life. They "have no desire to abandon any of their aims provided others have less means to further theirs" (TJ 144). I argued in Chapter II.2.4 that this stipulation of the original position is inconsistent with Rawls's adoption of the belief-desire model of motivation, because envy is an unavoidable consequence of being motivated by desire alone.

The parties in the original position are also, thirdly, capable of a sense of justice. This means that they will be able to respect and carry out whatever principles of justice are chosen, such that their strict compliance with such principles can be assured at the outset (TJ 145). The intuition behind this stipulation has two parts. First, the parties are here assumed be autonomous and responsible; that is, they are assumed to be capable of adhering to their agreements without bribes, coercion, or external promptings. But second, they are assumed to have the ability to identify the principles they have chosen to distribute available resources as just, *even though they did not come together with the conscious aim of achieving justice*.

It is important to the value-neutrality of the metaethics of *A Theory of Justice* – the attempt to derive a substantive normative theory from "commonly shared ... widely accepted but weak ... natural and plausible ... seem[ingly] innocuous or even trivial ... [metaethical] premises" (TJ 18) – that they did not. That is, it is important that they choose principles of distribution on purely self-interested grounds that, on Rawls's view, contain no normative ethical bias. If, on the contrary, they had gathered with a view to choosing principles of justice explicitly, they would have had to agree upon in what justice consists. Since this question would have required a collective and transpersonal answer, the stipulation of mutual disinterest would have been violated. Hence Rawls would have had to make the metaethically controversial – because normatively biased – motivational assumption that the parties were at least in part altruistic. This, in turn, would have undermined the strength of his eventually deductive argument, since it would have been less impressive to derive normative conclusions from explicitly normative premises than from arguably nonnormative ones. So to stipulate that the parties have a sense of justice is to ensure that they recognize it when they find it, even though they were not motivated to look for it. This stipulation enables Rawls to preserve at least the semblance of value-neutrality in his premises, while at the same time establish in his conclusion a normative motivational basis for citizens' adherence in the well-ordered society to the principles of justice that structure it.

Finally, the parties are limited in the information available to them for making the choice by what Rawls calls *the veil of ignorance* (TJ 136-142). They do not know their personal identities, the society to which they previously belonged, their socioeconomic status in it, their conceptions of the good, or any other particular facts that might bias their choice of principles. For example, if one knew one's particular society, one might tend to choose principles of justice that would improve it specifically; if one knew one's socioeconomic status in it, one might tend to choose principles that would enhance or protect it; if one knew one's conception of the good, one might tend to choose principles that would favor the particular distribution of resources it required (TJ 137). By excluding all such information, no individual can use these actual resources, status or circumstances as leverage, threat, bribe or coercion to fashion principles to his own advantage. However, the parties do know the general facts about human nature, the laws of the natural and social sciences, and any other general facts that enable them to choose principles of justice.

Together these constraints express the following assumptions. First, the choice of principles should not be biased by the personal advantage or disadvantage of the choosers. Second, it should not be possible to tailor principles of justice to one's own interests. And third, all choosers are equal in having a conception of the good and a sense of justice. Under these circumstances, the requirements of rational deliberation can be observed and unanimity can be insured:

> [S]ince the differences among the parties are unknown to them, and everyone is equally rational and similarly situated, each is convinced by the same arguments. Therefore, we can view the choice in the original position from the standpoint of one person selected at random. If anyone after due reflection prefers a conception of justice to another, then they all do, and a unanimous agreement can be reached (TJ 139).

Together with the circumstances of justice, the original position constitutes Rawls's version of the preconditioning circumstances that is traditionally invoked to metaethically justify the social contract.

3.2. Rawls's Normative Theory

3.2.1. The Two Principles of Justice

Rawls then argues that, under these preconditioning circumstances, the parties would choose his two principles of justice to structure and regulate their society. The first of these is what Rawls calls the *Millian Principle*. This says that each citizen of a well-ordered society has a right to the most extensive basic liberty compatible with similar liberty for others. All have equal liberties because all have equal rights under law designed to protect that liberty. The Millian Principle thus echoes and elaborates upon the

juridical ideal of equality characteristic of traditional Social Contract Theory. Rawls's second principle has two parts. Part one Rawls describes as the *Fair Opportunity Principle*. This requires that social and economic inequalities be attach to roles and offices available to all. Rawls calls the second part of the second principle the *Difference Principle*. This one requires that social and economic inequalities benefit the least advantaged (TJ 60).

Rawls is the first Social Contract Theorist to try to meet the Marxist objection that merely juridical equality is no equality at all. He does not do this by arguing in favor of perfect social and economic equality for all citizens. This would have certain undesirable consequences, such as undermining incentives to do any of those things that might be driven by the desire for individual social or economic advancement – difficult or challenging or committed work, for example; or innovation or invention, or activities or roles that require a high level of visibility and so a high level of risk. Instead the second principle, in both of its parts, requires that the rewards of any such advancement benefit even those who are least likely to obtain them. So, for example, the more economically advantaged may be taxed at a higher rate so as to subsidize education, housing or health care for the least advantaged; the more socially powerful may be required to exercise that power – through judicious governance, role modeling, charity, or advocacy of social causes – in ways that benefit the less powerful.[9]

Rawls argues, but does not claim that the parties would argue, in favor of a serial ordering of these principles, such that the Millian Principle has lexical priority over the second, the Fair Opportunity Principle has lexical priority over the Difference Principle, and so the Millian Principle has lexical priority over the Difference Principle (TJ 62-3). This means that it is impermissible to abdicate or restrict basic liberties in order to obtain greater social or economic benefits, and similarly impermissible to abdicate either basic liberties or one's right to consideration for any public role or office in order to obtain any such benefits (TJ 61). So, for example, it would be impermissible to pass laws permitting press censorship for the sake of greater social harmony or economic well-being among citizens. Of course this does not mean that there are no restrictions at all on basic freedoms. We must not be at liberty to shout

[9]In the Reagan and Bush eras, discussion of Rawls's Difference Principle found its way into rationalizations of "trickle-down economics," in which the deregulation of corporate investment practices was supposed to be justified by creating more jobs for the economically disadvantaged. This was a perversion of Rawls's conception of justice to begin with, and has proven to be false in its predictions as well. Nietzche would have been similarly horrified by the use made of his concept of the *Übermensch* by the Nazis, as Marx would have been by the atrocities committed by Lenin and Stalin in the name of the "truly human society." Whether a philosophical conception of the good society can ever have tangible social influence except through misunderstanding or misapplication of it is an issue I neglect with relief.

"Fire!" in a crowded theater just for fun, nor to allow children to act or be treated in any manner whatsoever. But in all such cases, basic liberties can be restricted only for the sake of greater liberty for everyone (TJ 244).

Similarly, an example of violating the lexical priority of the Fair Opportunity Principle over the Difference Principle would be any tradeoff that benefited the least advantaged economically by restricting or canceling their right to apply for public roles or offices. An hereditary but benevolent aristocracy, primogeniture, or laws benefiting needy groups or individuals with tax credits or bonuses or welfare payments in return for prohibiting them from running for election to public office would all violate the lexical priority of the Fair Opportunity Principle over the Difference Principle, by reducing the life prospects of those disadvantaged individuals (TJ 299-301). Finally, an example of violating the lexical priority of the Millian Principle over the Difference Principle would be laws or customs permitting the sale of oneself into slavery, or abdication of the right to vote, if it improved one's economic well-being to do either. Rawls thus contrasts what he calls his *special conception of justice* with a *general conception* that would permit the unequal distribution of any and all social and economic resources if so doing were to improve the situation of the least advantaged.

3.2.2. Primary Goods

Rawls's conception of the social and economic resources distributed by the principles of justice the parties in the original position decide upon is equally distinctive. Rawls defines *primary social goods* as those instrumental resources which are directly under the control of principles of distribution, such that one necessarily wants to maximize them in order to achieve whatever else one's goals may be (TJ 62, 92-93). They comprise, first, the rights and liberties distributed by the Millian Principle: the freedom to vote, freedom of speech and assembly, of conscience and thought, of one's person, to hold private property, and finally freedom from arbitrary arrest and seizure. Rawls argues that these are each necessary and jointly sufficient for a further primary good, that of self-respect. This he defines as a person's sense of her own value, the conviction that her goals are worth pursuing, and a realistic confidence in her ability to carry out her intentions (TJ Pars. 29 and 67).

The basic idea of self-respect as a primary social good is that a society that gives explicit priority to the freedom of its citizens from undue constraints, that explicitly values the pursuit of individual goals and interests as such, conveys to its citizens a sense that their goals and interests are worth pursuing. Such a society endorses and undergirds its citizens' aspirations to autonomous self-realization. The explicit juridical valuation of an individual's goals and interests as such will then confer on that individual a sense of value, or self-respect. Self-respect is thus a backward-looking consequence of a

socially endorsed respect for individual liberty: Implementation of the Millian Principle confers worth on individual goals, and these juridically valued goals in turn confer worth on the individual whose goals they are. Individuals then have self-respect because the fundamental principles that structure their society acknowledge and support their right and ability to pursue their ends as they please.

To see the force of this, consider two individuals who do not regard their goals as worth pursuing. For one, a realistic assessment of the limited social and economic resources available to him might require him to assign a much lower priority to his aspirations than to the exigencies of immediate survival. It might not be that he views the desire to become a lawyer as intrinsically worthless. Having had considerable contact with lawyers, he might think that lawyers are valuable, and that his skill in analytical reasoning would make him a good one. But if his first priority were to find a soup kitchen, his second to find a place to sleep for the night, his third to avoid being sent to jail on vagrancy charges, and his fourth to learn how to read, his dream of becoming a lawyer would not count as a goal at all, for it would remain for all intents and purposes merely an unrealistic fantasy. Since he could not be said even to have this as a goal in the absence of sufficient social and economic resources, he could not even be described as viewing his goal as worthless.

A second individual might regard her aspiration to become a lawyer as not worth pursuing because she regarded herself as lacking in worth. Although she might nurse the secret desire to be a lawyer, she would not be able to imagine herself doing the job well, and would lack the self-confidence necessary to do what was instrumentally necessary to become one. The thoughts of applying to law school, taking the LSATs, clerking for a judge all would make her feel afraid, and sure she would fail. The basic idea behind Rawls's conception of the basic liberties as necessary and conjointly sufficient for the primary good of self-respect is that the citizen of a well-ordered society structured by Rawls's two principles of justice will not experience such failures of nerve. A citizen whose society publicly endorses freedom of thought, speech and mobility thereby publicly endorses the free pursuit of her individual goals as valuable simply because they are her goals. This public valuation of her goals as hers will lead her to regard herself as similarly valuable. This, in turn, will give her the self-confidence to pursue them.

Other primary goods distributed by the two principles of justice include the powers and opportunities distributed by the Fair Opportunity Principle, and income and wealth distributed by the Difference Principle. Both principles require that any disparities in the distribution of these goods be to the benefit of the least advantaged, i.e. such that any advantages accruing to those who are better off also work to the benefit of those who are less well off, and any decrease in those advantages would thereby decrease the benefits to the less well off. So, for example, a thriving business with a strong profit

margin arguably creates more jobs for those who need them through reinvestment and expansion; the lower the profit margin, the less revenue is available to salary positions, and so the fewer jobs available and the worse off the job seekers become. Both parts of the second principle of justice thus satisfy the requirement of Pareto optimality, i.e. that the preferred distribution is one that cannot be improved so as to increase anyone's benefits without cost to someone else (TJ 79). Taken together, rights, liberties, powers, opportunities, income, wealth and self-respect constitute resources which, on Rawls's view, the parties in the original position would want to maximize for themselves as necessary means to the achievement of whatever conception of the good they might hold.

Rawls's conception of primary social goods as necessary social and economic resources distributable by principles of justice circumvents a problem that besets Utilitarian theories of justice (TJ 91-92). On all such views, justice consists in maximizing the total or average utility, summed over all citizens, where "utility" is a dominant final end defined in various ways depending on the specifics of the theory: as happiness, welfare, desire-satisfaction, or pleasure (I consider the classical version of Utilitarianism in Chapter XII). All of these subjective states of the agent raise the problem of how to measure the wellbeing of one citizen relative to another, i.e. of how to make interpersonal comparisons of utility. This problem was addressed in depth in Chapter IV.1. Suffice it here to say that without being able to make these interpersonal comparisons of utility, no variety of Utilitarianism has a basis on which to apply its distributive principle of justice, for there is no way of ascertaining whether I am in objective fact happier than, just as happy as, or less happy than you; nor, therefore, whether, if I am less happy than you, I am nevertheless as happy as I can be. Without ascertaining these things, there is no way of knowing whether or when total or average happiness, welfare, desire-satisfaction, or pleasure has been collectively maximized. I examine Brandt's solution to this question in Chapter XI.

Rawls's concept of primary social goods bypasses it entirely. First, primary goods are publicly quantifiable. Even self-respect, the most elusive of the set, is tied systematically to rights and liberties that are less elusive and so more susceptible to comparative measurement: One individual has less self-respect than another if the well-ordered society defends less ardently the rights and liberties of the former than it does those of the latter. Second, they are stipulated to be consensually valuable regardless of one's conception of the good. In this they diverge from Utilitarianism, which requires all citizens to adopt the same ultimate, dominant-end conception of the good. Third, the lexical ordering of Rawls's principles of justice assigns to the seven kinds of primary goods a ranking that is both objective and ordinal, and requires only that each be maximized within its ordinal position for all citizens so far as

possible. Since each is publicly quantifiable at least in general terms, there is far less difficulty in saying in what maximization of a primary good consists.

In outline, then, Rawls's theory of justice has the following structure:

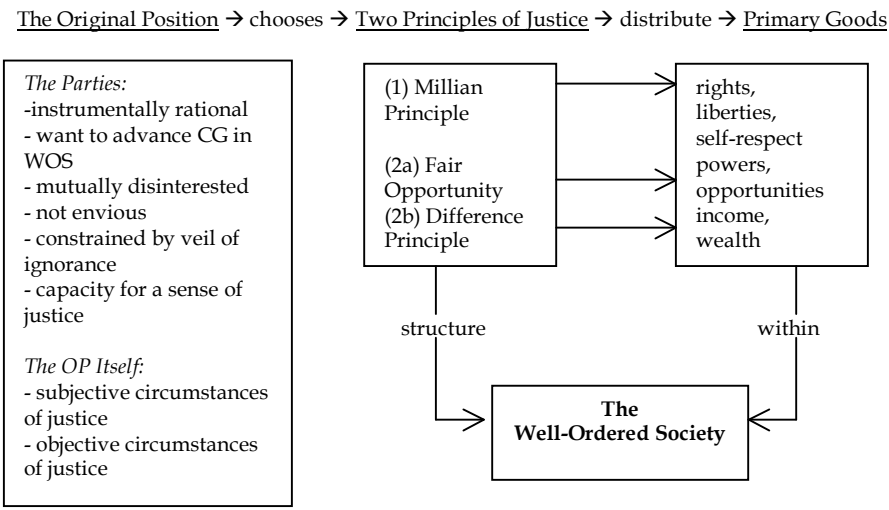


Figure 13. The Structure of Rawls's Theory of Justice

---

4. Habermas's Critique

4.1. Primary Goods

Jürgen Habermas objects to Rawls's normative characterization of rights and liberties as primary goods on the grounds that, first, rights and liberties are not things that can be possessed in the ways that income, wealth, or property can. Rather, they regulate interactions among agents. Second, they must be exercised, not merely owned, in order to be enjoyed (RT 54).[10]

> Only between rights, on the one side, and actual chances to exercise rights, on the other, can there exist a chasm that is problematic from the

[10] "Reconciliation Through the Public Use of Reason: Remarks on John Rawls's Political Liberalism," *The Journal of Philosophy XCII*, 3 (March 1995), 109-131; reprinted in *The Inclusion of the Other: Studies in Political Theory*, trans. Ciaran Cronin (Cambridge, Mass.: MIT Press, 1998). Page references to the latter edition are paginated in the text in brackets and preceded by "RT".

> perspective of justice; such a rupture cannot exist between the possession and enjoyment of goods (56).

Habermas thus offers a counterpart to Marx's critique of traditional Social Contract Theory, that rights are empty, merely juridical concepts without the fair distribution of economic resources that gives them meaning. Similarly, Habermas faults Rawls for reifying juridical concepts that are just as empty and meaningless without practical and concrete opportunities for their exercise. As is true for Marx, such opportunities presuppose the material and social means to do so. In this respect there is an asymmetry between rights and goods that, Habermas argues, Rawls fails to acknowledge.

Furthermore, Habermas reproaches Rawls at the metaethical level, for having effectively conflated the distinction between teleological values and deontological norms by conceiving of rights and liberties as goods. Norms, Habermas argues, guide action decisions, impose obligations, express behavioral expectations, make binary claims of validity or invalidity, are impartial in their application to agents, and must be mutually coherent. Values, by contrast, guide outcome choices, offer arrays of particular objects of preference, express purposive ends, require ordinal rankings, are culturally or subjectively conditioned, and compete for priority.

> To sum up, norms differ from values, first, in their relation to rule-governed as opposed to purposive action; second, in a binary as opposed to a gradual coding of the respective validity claims; third, in their absolute as opposed to relative bindingness; and last, in the criteria that systems of norms as opposed to systems of values must satisfy (55).

Rawls, Habermas thinks, ignores these distinctions by assimilating rights and liberties into the decision theorist's pairwise comparisons among preference alternatives.

However, from Rawls's stipulation of rights and liberties as objects of preference along with other goods, it does not follow that he disregards these deep structural differences between them, any more than it would follow from my having to choose between darning a sock and reading a book that I disregarded the structural differences between them. We have already seen in Chapter IV that it is both a blessing and a curse that canonical decision theory can reify any state of affairs that can be an object of desire into a preference alternative. And surely rights and liberties can be objects of desire. Habermas is right to call attention to the important difference between those primary goods which can be enjoyed merely by being owned and those which can be enjoyed only by being exercised. This distinction assumes increasing significance as the primary goods are gradually distributed by the two principles of justice under the four-stage sequence. But we can see in Figure 12 above that Rawls does, in fact, make provision for this distinction in the primary goods that each principle of justice distributes: rights and liberties are distributed specifically by the first, Millian principle that stipulates that the

distribution of these goods must be equal. This stipulation in effect ensures that the particular, norm-governed character of rights and liberties that Habermas enumerates – rule-governedness, binariness, unconditionality, and mutual coherence – are respected.

4.2. The Four-Stage Sequence

Once the principles for just distribution of the primary goods are chosen, Rawls stipulates a process, which he calls *the four-stage sequence*, during which progressively more information is made accessible to the parties in order that they may make increasingly specific choices about constitutional and legislative matters (TJ 195-201). As the veil of ignorance is gradually lifted, the parties learn increasingly more about their actual places in society and the particular distribution of primary goods their choice of the two principles has allotted them.

Here Habermas observes that first, in order to ensure the consonance of the two principles of justice with the more specific information that comes in as the veil of ignorance is lifted, the designer of this normative moral theory must anticipate and foresee the contents and effects of all the specific information he denied to the parties in the Original Position in the first place (RT 58). This authorial omniscience, assuming it is plausible, undermines the function and purpose of the veil of ignorance itself, and leaves the parties in the original position with virtually nothing to do. Second, similarly,

> the higher the veil of ignorance is raised and the more Rawls's citizens themselves take on real flesh and blood, the more deeply they find themselves subject to principles and norms that have been anticipated in theory and have already become institutionalized beyond their control. In this way, the theory deprives the citizens of too many of the insights that they would have to assimilate anew in each generation (RT 69).

Because all of the substantive questions of just distribution have been decided before the citizens find out who they are, either by the parties in the original position under the veil of ignorance, or by the author who works out the detailed social arrangements that the veil of ignorance temporarily conceals, there is virtually nothing for the citizens to do either, beyond passively accepting and affirming (Rawls's term) the principles and arrangements that have been fashioned for them. This ensures social stability, but only at the cost of citizens' political autonomy and their active exercise of public reason (RT 70). The four-stage sequence, Habermas objects, takes the important work of theory-building and political decision-making out of the hands of the citizens whose society is their subject, leaving them with little to do but enjoy the fruits of intellectual and political work performed at earlier stages of conception; and, therefore, no resources for training future generations to do this important work. Rather, the citizens are better understood as custodians of the well-ordered society whose main responsibility is to ensure its stability

and continuity over time. But because they are deprived of the opportunity to engage actively in the project of determining the general content and particular form of the principles of justice that fix the basic structure of their society, the validity of those principles must remain in question.

4.3. The Moral Point of View

Habermas's most serious criticism is that this lack, in turn, deprives the citizens of the well-ordered society of a genuinely moral point of view, in which

> everyone is required to take the perspective of everyone else and thus to project herself into the understandings of self and world of all others; from this interlocking of perspectives there emerges an ideally extended 'we-perspective' from which all can test in common whether they wish to make a controversial norm the basis of their shared practice (RT 58; cf. also 68, 81-82) [.]

Habermas's conception of the moral point of view thus requires a version of Mead's ideal role-taking,[11] in which we enlarge our perspectives beyond the personal, by successively assuming the personal perspectives of all other participants and incorporating them into our own. The ideal end-point of this process is a "universally valid view of the world," in which "what from my point of view is equally good for all [would] actually be in the equal interest of each individual" (RT 57). But the process as Habermas and Mead conceive it require a degree of imagination and insight into the inner lives of others that can come only from extensive experience of and intensive information-gathering about different kinds of people and cultures. Thus it is strongly conditioned by the empirical experience of walking in another's shoes: living in another culture, for example, or visiting a different political milieu, or taking a job outside one's class and education status, or inhabiting for an extended period of time a different social environment, or fraternizing with friends from other subcultures.

There is no question that concerted use of the imagination, combined with sensitivity to others' attitudes, curiosity about their origins, and a consequent openness to their contributions to serious discussion can have the similarly salutary effect of opening one's eyes, deepening one's insight into otherness, and raising one's awareness of the subjectivity of the personal values and attitudes with which one began. However, Habermas's optimistic assumption that this process will engender a perspective from which the successive perspectives one has assumed can be unified in a universally valid "we-perspective" holds only in cases in which the perspectives in question are not in fact structurally irreconcilable.

---

[11] See George Herbert Mead, "Fragments on Ethics," in *Mind, Self and Society* (Chicago: University of Chicago Press, 1934), 379 ff.

But Rawls is concerned precisely with such cases, and builds them into his conception of the original position; for these are the subjective and objective circumstances of justice under which truly urgent questions of political justice arise.[12] In such cases of incorrigible conflict among personal perspectives, the process of ideal role-taking must engender not a point of view from which the interests of all can be equally served; but rather the recognition that the interests of some must be sacrificed. This point of view is the perspective of transpersonal rationality that prepares one to accept the possible sacrifice of one's personal interests in the service of a larger vision of moral and social justice. These are among the possibilities that the parties in the original position must consider. By operating under the veil of ignorance, the parties accept that "any principle chosen in the original position may require a large sacrifice for some" (TJ 176). No participant in rational dialogue who is unprepared to countenance such a possibility, and to accept it as a possible outcome, can be said to have a genuinely moral point of view toward the resolution of questions of just distribution.

Habermas's objection that Rawls's citizens of the well-ordered society lack such a point of view is therefore misplaced on two counts. First, these sacrifices have already been made in the original position, through imposition of the veil of ignorance; and resolution of questions of just distribution already achieved. The citizens of the well-ordered society are the result of such a resolution, not the agents of it, by definition. Hence the transpersonally rational perspective that is a precondition for making such sacrifices is no longer required in the well-ordered society. But second, this perspective in fact is, *contra* Habermas, available to its citizens, through reflection on the concept of the original position and the principles of justice it engenders. Here it is worth quoting Rawls's concluding description of this standpoint at length:

> Each aspect of the original position can be given a supporting explanation. Thus what we are doing is to combine into one conception the totality of conditions that we are ready upon due reflection to recognize as reasonable in our conduct with regard to one another. *Once we grasp this conception, we can at any time look at the social world from the required point of view.* … it is a certain form of thought and feeling that rational persons can adopt within the world. And having done so, they

[12] At least in the United States. I suspect that this difference in what Rawls and Habermas respectively require for the moral point of view is at least in part a function of their differing background political environments: the United States is a four hundred year-old attempt to contain murderously antagonistic and widely diverse social and cultural forces under the rubric of capitalism, whereas Germany is a sixty-year-old attempt to contain relatively homogeneous social and cultural forces under the rubric of democracy.

> can, whatever their generation, *bring together into one scheme all individual perspectives and arrive together at regulative principles that can be affirmed by everyone as he lives by them, each from his own standpoint.* Purity of heart, if one could attain it, would be to see clearly and to act with grace and self-command from this point of view (TJ 587; italics added).

Thus both Rawls and Habermas agree that the moral point of view integrates all individual perspectives into a single transpersonal one. But Rawls's conception of the moral point of view is one that results from agreement reached in the original position. It is accessible both to us as readers and to the citizens of Rawls's well-ordered society, whenever we need to be reminded of the magnitude of the efforts that had to have been made so that justice could be achieved. Embedded in the conditions and outcome of the original position, it provides us and the citizens of the well-ordered society with a mnemonic device for recalling and affirming the validity of the principles of justice that have been formulated to structure it. Because it presupposes prior resolution of conflicting claims to scarce resources, it is a much more demanding point of view than Habermas', for its vision requires a wider sweep and greater distance from the pull of any particular agent's self-interest.

By contrast, Habermas conceives the "we-perspective" not as a consequence but rather as a precondition for reaching agreement on principles of justice. This requires that conflicting claims among individuals be reconcilable in advance of such agreement – i.e. that differences among individual perspectives not be so deep that they cannot be coherently integrated *independently of and prior to any accord on principles of justice that are supposed to resolve them*. However, to meet this requirement is, in effect, to leave such individuals with nothing substantial to resolve.

So it is Habermas's conception of the moral point of view, not Rawls's, which leaves participants in rational dialogue with nothing to do. For Habermas's conception deprives citizens of the painful and instructive opportunity to engage the difficult issues of whose interests are to be advanced and whose sacrificed, whose claims to scarce resources are valid and whose invalid, and what principles of justice are most suitable for all regardless of personal advantage to anyone. A moral point of view that successfully integrates all individual perspectives into a single transpersonal one presupposes that these issues have been addressed and resolved. It can therefore be a consequence of such resolution but not a precondition of it.

5. The Analogy with Science Reconsidered

5.1. Pure Procedural Justice

We have seen that the parties in the original position are stipulated to have the capacity to recognize the two principles of justice as, indeed, just. But

in what sense are these principles just? Rawls distinguishes three ways of conceiving justice. The first is what he calls *perfect procedural justice* (TJ 85). Here we begin with a prior conception of what justice under the circumstances requires, and then devise a procedure for yielding that outcome. The example he gives is of dividing a tasty cake fairly among healthy adults. Intuitively, it seems obvious that each should have a slice equal to that of everyone else's. To insure this we ask the person who will get the last slice to cut the cake for everyone. Most problems of fair distribution are not solved so easily.

In *imperfect procedural justice*, by contrast, we again have an independent conception of what justice requires, but no sure procedure for reaching that outcome. Here Rawls's example is of a criminal trial. Our prior conception of what justice requires is simply that the defendant should be found guilty if and only if he committed the crime. We do our best to secure that outcome by devising and following laws of evidence, testimony and deliberation, but it does not always work; there are miscarriages of justice. Classical Utilitarianism would provide another example of imperfect procedural justice (TJ 89): It begins with a prior conception of what justice requires – i.e. the greatest net sum of happiness, welfare, or desire-satisfaction for everyone, and may experiment with different procedures – for example, regulated or unregulated free markets, centrally planned economies, private ownership versus nationalization of industries, employee stock ownership or profit-sharing plans, etc. – for maximizing it. But it can provide no procedure guaranteed to yield this outcome.

Rawls's conception of justice as fairness is an example of what he calls *pure procedural justice*. Here there is no prior conception of what justice requires, independent of a certain kind of procedure, which, when actually carried out, ensures its outcome as just by definition. Therefore, pure procedural justice generates a just outcome from a value-neutral starting point, in that the procedure that generates this outcome can be specified as a series of steps to be followed and conditions to be met that include no reference to any independent preconceptions about in what justice should consist. The example he gives is of gambling: If all players voluntarily engage in a series of unrigged bets and no one cheats, such that "unrigged" and "cheats" are shorthand for subprocedures that can be specified value-neutrally, the resulting distribution of cash is by definition fair (TJ 86), and recognizable as such by all players. Similarly, Rawls argues, if the parties in the original position go through the process of deliberation and agreement he describes under the conditions he describes, the outcome is similarly fair by definition and recognizable as such by them.

In both cases, the procedure actually must be carried out in order to judge the resulting outcome as just, and in order to know what justice requires under those circumstances. Unless it is the actual result of that very

procedure, the particular distribution has no claim to justice. So, for example, there is nothing inherently just about your having 86% of the pot compared to my 2%, unless this is the outcome of a gamble in which we both voluntarily participated. Only its status as the outcome of that procedure makes it just. The way in which the two principles of justice distribute primary goods in the well-ordered society is just, on this view, because and only because of the procedure of deliberation and agreement in the original position through which they have been selected. However, Rawls's integration of widely accepted principles of rationality into the conception of the original position insures that the resulting distribution will not seem arbitrary.

Rawls's account of pure procedural justice develops further the conception of moral objectivity he limned in "Outline of a Decision Procedure for Ethics," discussed in Section 1. Like that account of moral objectivity, pure procedural justice might be compared to an analogous conception in the natural sciences, which we might call *pure procedural truth*. Like pure procedural justice, pure procedural truth would presuppose no criterion of validity independent of the actual outcome of a correctly carried-out procedure for ascertaining it. Therefore, like pure procedural justice, pure procedural truth would be value-neutral in the important sense that it would be unbiased by any preconceptions about what a legitimate outcome ought to look like. More specifically, pure procedural scientific truth, on this hypothesis, would be defined as the outcome – whatever that outcome might be – of correctly carried out procedures of observation, data gathering, inductive reasoning, hypothesis construction, deductive reasoning and prediction, experimentation under controlled conditions, and intersubjective replication of predicted experimental results. Whatever outcomes resulted from this procedure would qualify as by definition objectively valid.

Now natural scientists do not actually adhere to this conception of pure procedural truth in all of its particulars, nor would they necessarily accept the outcomes of these procedures as such without qualification. On the contrary: we know that scientists very often proceed unsystematically and intuitively, follow hunches rather than rational deliberation and deviate from canonical scientific procedure at many points along the way. Moreover, even the more scrupulously procedural may shrink from anointing the outcome of their labors with the appellation of objective validity. For suppose that outcome is too anomalous relative to their initial premises, or deviates too far from accepted scientific dogma of the day, or violates too radically the personal metaphysical views that underlie their acceptance of that dogma. Then rather than legitimate it as objectively valid, no matter how many times that outcome has been replicated under controlled conditions by different laboratories, they may blink at the requisite inference rather than stare it down. They may infer instead that mistakes in procedure must have been made somewhere in order to generate this outcome; or suspect that evidence

has been doctored; or turn their attention to all the additional errors that can accumulate in the process of intersubjective replication of results by different laboratories. That is, they may, in effect, conceive objective validity as *imperfect procedural truth*, in which one begins with a prior conception of scientific truth, for the discovery of which one has devised rather vulnerable procedures of reasoning and experimentation that do not always work. If the outcome of these procedures in a particular case deviates too wildly from one's preconceptions about what an objectively valid outcome should be, the inference is available that it is the procedures, rather than one's preconceptions, that are at fault.

5.2. Wide Reflective Equilibrium

Rawls's early analogy with scientific procedure, carried through here, illuminates the question of which account of justice – pure procedural, perfect procedural, or imperfect procedural – in fact best fits Rawls's account in *A Theory of Justice* of the eventually deductive relationship between the original position and the well-ordered society as structured by the two principles of justice. Like experimental procedure among scientists, deliberative procedure among the parties is stipulated to generate an outcome that is objectively valid in virtue of its history. Like an experimental result, the deliberative result may cast doubt on the care or precision with which the procedure was executed. Or, in case these standards have been verifiably met, it may reveal that our commitment to the procedure as an index of objective validity is in fact outweighed by our commitment to an independent preconception about what an objectively valid outcome can be – to which this result has perhaps failed to measure up. Is the relationship between the original position procedure and the two principles of justice more like the procedure and results of gambling? Or is it more like the procedure and results of a criminal trial?

So far we have examined Rawls's metaethics from the perspective of the original position. We have traced just a few of the basic relationships between the conditions that define it and the principles it is claimed deductively to generate. What we have not yet done is the analogue of what scientists do in assessing the status of their experimental results relative to the hypotheses that predicted them, namely decide whether they are in accordance with their trained judgment, intuitions, and practice as scientists; or whether those results, the procedures that generated them, the hypotheses those procedures were intended to test, or indeed the training and practice on which all depend require rethinking or revision.

Analogously, we have not yet assessed whether the metaethical and normative scheme Rawls offers conforms to our commonsense moral intuitions about what a free, equal and impartial consensus in deliberation entails, or in what a just distribution of moderately scarce social goods in a

well-ordered society consists. Nor, if either the metaethical or the normative scheme does not, whether it is our commonsense moral intuitions, the description of the original position, or the statement of the two principles of justice that should be rethought. In later sections, I show that we need to be able to make these assessments of Rawls's scheme, independently of the Instrumentalist strategy by which he first attempts to justify it; and that his later revisions of this strategy retain its deductive character while jettisoning its Instrumentalism (recall that Instrumentalism is just one kind of Deductivism).

Rawls's concept of wide reflective equilibrium allows us to make these assessments. He distinguishes three elements in the justification of his theory of justice. The first would consist in showing that the original position expresses shared, weak assumptions that we accept about the conditions of rationality and impartiality under which principles of justice should be chosen (TJ 19). The second would show that the two principles of justice derived from these conditions express or extend our basic moral intuitions about a just society, or enable us to resolve cases in which our intuitions are unsure (TJ 21). The third combines the first two: We try to achieve *narrow reflective equilibrium* between the first two elements – the original position and the two principles of justice – by checking each against the other in relation to the third: our basic moral intuitions. We examine or modify the statement of the two principles in relation to the construction of the original position that generated them, with an eye to maximizing their coherence with one another, as well as with our commonsense moral intuitions. By aiming to maximize the coherence of all three elements – the description of the original position, the statement of the two principles of justice, and our commonsense moral intuitions, we enter into a process of self-scrutiny in which we gradually and reflectively articulate considered moral judgments (TJ 47-48), leaving open the possibility that any of the three may undergo radical revision in the process.

We then try to achieve *wide reflective equilibrium* by testing and revising these considered judgments in light of hard cases that may challenge them and other theories of justice – Classical Utilitarianism most importantly, for Rawls, but others such as Marxism and Perfectionism as well – that offer competing principles of distribution (TJ 49-50). Rawls's overall scheme in wide reflective equilibrium would then look this way (note the bi-directionality of some arrows of influence):

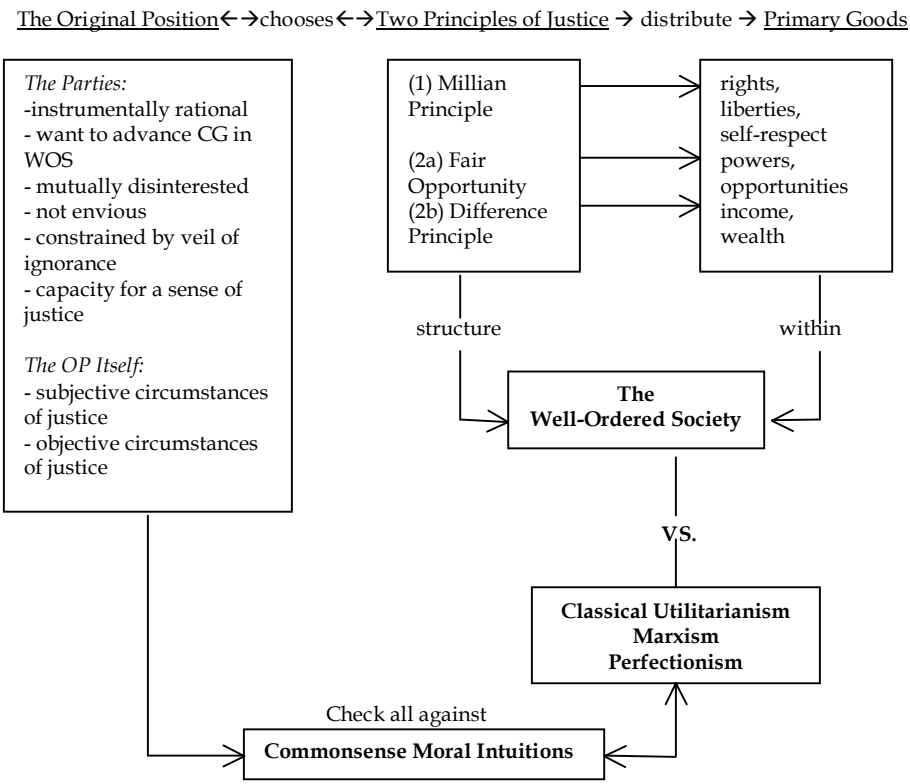


Figure 14. Wide Reflective Equilibrium

The analogy with scientific procedure enables us to clarify what is at stake. In the process of constructing a scientific theory, scientists are influenced by competing hypotheses and explanations to make adjustments and revisions both to the statement of their own and to the predictions they expect it to yield. But both hypothesis and predictions are also conditioned by a scientist's epistemic intuitions about what is physically possible and likely under the circumstances. If a hypothesis entails predictions that are, in that scientist's judgment, highly unlikely to occur, this undermines the plausibility of the hypothesis in favor of competing ones, before the experiments have even been performed.

On the other hand, such judgments of plausibility may have to be sacrificed if the hypothesis has been carefully formulated on the basis of sufficient inductive evidence and rules of inference scrupulously observed. In this case, the strangeness or improbability of the predicted results may not warrant further revision or rejection of the hypothesis before experiments have been performed. Should the results of such experiments then be equally strange or improbable, a commitment to the rational integrity of the procedure necessitates sacrifice of those epistemic intuitions, not an insistence that its integrity must have been violated somewhere along the line.

Thus in the scientific case, an overriding commitment to the rationality of scientific procedure requires conceiving of it as an instance of pure procedural truth; whereas an overriding commitment to one's prior conception of what might plausibly constitute scientific truth – a commitment that might, in a particular case, require rethinking or revising the procedure in order to conform to it – requires conceiving of this procedure as an instance of perfect procedural truth. The question is, to paraphrase Humpty Dumpty, which – scientific procedure or conception of scientific truth – is to prevail.

Analogously in Rawls's theory of justice. Which – deliberative procedure or commonsense moral intuitions – is to prevail? If our commonsense moral intuitions about what justice requires are the final arbiter in evaluating the plausibility of the two principles, or of their derivation from the original position, or of the original position itself, such that we are unwilling to sacrifice these intuitions to any counterintuitive results Rawls's stipulated decision procedure (correctly carried out) might have, then Rawls's deliberative procedure is in fact one of perfect procedural justice. For by insisting that the outcome of this procedure finally conform to our commonsense moral intuitions about what justice requires, we will have subordinated that outcome, and indeed the procedure itself, to prior preconceptions about what justice requires that the procedure is designed to approximate. Only if our moral intuitions are equally vulnerable to sacrifice in the service of the rational procedure of deliberation Rawls describes can he make the claim about justice analogously to that one might wish to make about science. Only if rationality may outweigh moral intuition in metaethics just as it may outweigh epistemic intuition in science is Rawls's view truly an instance of pure procedural justice; and so only then can he claim to have met the standard of moral objectivity he set for himself in 1951. We will be in a better position to settle this question after we examine more closely what Rawls's deliberative procedure entails.

6. The Continuity Thesis

With at least some of the basics of Rawls's theory in place, let us now look at the implications of a certain metaphysical thesis the truth of which is presupposed in various early objections that were raised against *A Theory of*

Justice by a number of philosophers. Although their criticisms differ in many respects, all of these objections concur in assuming what I refer to as the *continuity thesis*. This consists of the following claims conjointly:

> (1) The parties in the original position are, and know themselves to be, fully mature persons who will be among the members of the well-ordered society which is generated by their choice of principles of justice.
>
> (2) The original position is a conscious event among others, integrated (compatibly with the constraints on knowledge and motivation imposed on the parties) into the regular continuity of experience that comprises each of their ongoing conscious lives.
>
> (3) The parties in the original position thus are, and regard themselves as, *psychologically continuing persons*, partially determined in personality and interests by prior experiences, capable of recollection and regret concerning the past, anticipation and apprehensiveness regarding the future, etc.

Thus, for example, some early criticisms of Rawls's *Theory of Justice* centered on what they took to be the individualistic assumptions embodied in the original position: Adina Schwartz argued that Rawls's assumption that the parties prefer a greater rather than a lesser amount of primary goods would contribute to a well-ordered society based on a preference for more rather than less wealth, and that this condition would be unacceptable to one who discovered herself to be a socialist.[13] Similarly, Thomas Nagel argued that the very concept of primary goods biases the choice of principles individualistically, against conceptions of the good that depend on the social interrelationships among individuals, and so may require the parties in the original position to commit themselves to a set of social arrangements that contravene their deepest convictions once the veil of ignorance is lifted.[14] David Gauthier attacked Rawls's assumption of economic rationality, showing that parties guided by instrumental reasoning in the OP would choose, not principles to structure a society based on justice as fairness, but instead those that would structure a "private society" instrumental to the pursuit of their individual utility-maximization.[15] Finally, Richard Miller argued that an individual in the original position who turned out to have been a member of the ruling class with an acute need for wealth and power in

---

[13]Adina Schwartz, "Moral Neutrality and Primary Goods," *Ethics 83* (1973), 294-307. See especially pp. 304-6.

[14]In "Rawls on Justice," *The Philosophical Review 87*, 2 (April 1973), 220-34; reprinted in *Reading Rawls*, Ed. Norman Daniels (New York: Basic Books, Inc., 1974).

[15]David Gauthier, "Justice and Natural Endowment: Toward a Critique of Rawls's Ideological Framework," *Social Theory and Practice 3* (1975), 3-26.

the society preceding the original position would find her interests frustrated by the egalitarian requirements of the difference principle.[16] Each of these criticisms called attention to the possibility of a disparity between the interests or beliefs of the parties in the original position and the conditions they may confront in the well-ordered society that is supposed to result from their choice. Hence each presupposes the continuity thesis. Subsequent criticisms of Rawls's theory presupposed it as well.[17]

Although the continuity thesis as stated above is not at odds with any of the conditions that define the original position, its exegetical validity is a matter for discussion. I argue here that if it is indeed contained in or a consequence of Rawls's theory, then it reinforces Rawls's reliance on an Instrumentalist metaethical strategy. This then casts into doubt the capacity of the original position to generate or justify any principles of justice at all. On the other hand, if the continuity thesis is viewed as dispensable and unnecessary to Rawls's theory, then Rawls is correct in maintaining the irrelevance of the question of personal identity to the construction of his moral theory[18]. In this case, the Instrumentalist justification for the two principles of justice should be supplanted by a modified conception of wide reflective equilibrium. The considerations that form the bulk of this discussion thus provide a philosophical rationale for Rawls's recent revisions in the model of justification on which his theory of justice rests, and for his increasing emphasis on us as moral mediators between the original position and the well-ordered society.[19]

---

[16]Richard Miller, "Rawls and Marxism," in Daniels (*op. cit.* Note 14.).

[17]See, for example, Anthony Kronman's and Samuel Scheffler's comments on Rawls's Tanner Lecture, "The Basic Liberties and Their Priority," *The Tanner Lectures on Human Values, Vol. III* (Salt Lake City: The University of Utah Press, 1982).

[18]John Rawls, "The Independence of Moral Theory," *Proceedings of the American Philosophical Association 1975* (Presidential Address).

[19]See in particular Lectures I and III of his Dewey Lectures, "Kantian Constructivism in Moral Theory: The Dewey Lectures 1980," *The Journal of Philosophy LXXVII*, 9 (September 1980); "Justice as Fairness: Political not Metaphysical," *Philosophy and Public Affairs 14*, 3 (Summer 1985), 223-251; and his emphasis on the original position as a "device of representation .. set up by you and me in working out justice as fairness ...." in *Political Liberalism* (*op. cit.* Note 3), 22-28. In my September 1976 paper, "Continuing Persons and the Original Position," for Rawls's moral and political philosophy seminar, I suggested that Rawls reconceive the original position as a device applied to the "circumstances of moral conflict that regularly confront us .. [that] both insures the impartiality of our moral judgments and also yields substantive moral principles in accordance with which we can judge these issues;" and also that he accord greater emphasis to "ourselves as moral mediators between the original position and the well-ordered society" – i.e. you and me – as practitioners of wide reflective equilibrium. I repeated these suggestions in revising this paper for publication as "Personal Continuity and Instrumental Rationality in Rawls's Theory of Justice," *Social Theory and*

Now let us consider whether or not, given the textual evidence, anything like the continuity thesis is stated or implied by Rawls; and what problems for his theory, if any, turn on a positive or negative answer to this question. To begin with, there is much in *A Theory of Justice* to lend support to the continuity thesis. Certain passages on what Rawls calls the strains of commitment suggest that the parties in the original position are psychologically continuous with identifiable members of the well-ordered society ((1) and (2) of the continuity thesis). For example, when Rawls stipulates that the parties have a sense of justice in that they "can rely on each other to understand and act in accordance with whatever principles are finally agreed to. Once principles are acknowledged the parties can depend on one another to conform to them" (TJ 145), the importance of insuring that *these individuals* are, in the well-ordered society, capable of adhering to the commitment they made in the original position is evident. This is reemphasized later when Rawls asserts that

> In view of the serious nature of the possible consequences [of the original agreement], the question of the burden of commitment is especially acute. A person is choosing once and for all the standards which are to govern his life prospects ... the parties must weight with care whether they will be able to stick by their commitment in all circumstances (TJ 176).

These claims clearly presuppose that the parties in the original position are, and know themselves to be, psychologically continuous with particular members of the society the basic structure upon which they now decide. Also, in discussing sound procedures of moral education in the well-ordered society, Rawls proposes that "in agreeing to principles of right the parties in the original position at the same time consent to the arrangements necessary to make these principles effective in their conduct" (TJ 515). This condition is clearly meant to insure the conformity to principle in the well-ordered society of the parties in the original position. Finally, Rawls made clear in subsequent discussion[20] that the parties in the original position are to be conceived as future members of a well-ordered society.

Also relevant to the continuity thesis are those passages in *A Theory of Justice* which suggest that the parties in the OP are continuing persons in that they are partially determined in their tastes and values by events prior to the original position ((2) and (3) of the continuity thesis), and that this must be considered in the subsequent well-ordered society. Rawls claims, for example, that the parties in the original position are to decide in advance the principles

---

Practice 13, 1 (Spring 1987), 49-76, from which this chapter originates. Also see Section 8, below.

[20]"Reply to Alexander and Musgrave," *Quarterly Journal of Economics 88* (November 1974), 633-39.

which are to regulate their interaction (TJ 11, 31), and further that they do this without a knowledge of their more particular ends:

> They implicitly agree, therefore, to conform their conceptions of their good to what the principles of justice require, or at least not to press claims which directly violate them. An individual who find that he enjoys seeing others in positions of lesser liberty understands that he has no claim whatever to this enjoyment (TJ 31).

The evidence here is strong that Rawls is canvassing the possibility that one such individual might discover, when the veil of ignorance is lifted, that his prior personal interests conflict with the principle he has chosen in the original position. Rawls's response is that such interests are simply to be disregarded. Further, his controversial claim that "it may turn out, once the veil of ignorance is removed, that some of [the parties in the original position] for religious or other reasons may not, in fact, want more of these [primary] goods" (TJ 142) provides additional support for the thesis that the parties in the original position are continuing persons, partially determined by their psychological histories, for whom the original position is an event among others in their conscious lives. This is because the implication here, as in the passage quoted above, is that the parties in the original position might subsequently discover in themselves psychological tendencies or desires that are in no sense determined by the decision made in the original position, hence must be determined by forces prior to that event. Nevertheless, those forces must continue to operate after it in order for the requisite discovery to be made. Again, the same point is made even more strongly later:

> How can the parties possibly know, or be sufficiently sure, that they can keep such an agreement? ... any principle chosen in the original position may require a large sacrifice for some. The beneficiaries of clearly unjust institutions (those founded on principles which have no claim to acceptance) may find it hard to reconcile themselves to the changes that will have to be made. But in this case they will know that they could not have maintained their position anyway (TJ 176).

Finally, there are auxiliary passages which, when taken together, clearly buttress the conception of the continuity of the parties as identifiable individuals both prior and subsequent to their participation in the original position. On page 166 of *A Theory of Justice*, Rawls observes that "The persons in the original position know that they already hold a place in some particular society;" and later, in discussing the strategic advantages of the four-stage sequence, he says, "So far I have supposed that once the principles of justice are chosen the parties return to their place in society and henceforth judge their claims on the social system by these principles" (TJ 196). These two passages establish that the original position as an event is integrated into the continuing personal histories of the parties ((2) of the continuity thesis).

Further evidence of the continuity thesis is to be gleaned from the concluding paragraphs of the *Dewey Lectures*[21], where Rawls claims of the parties in the original position that

> persons so conceived and moved by their highest-order interests are themselves, in their rationally autonomous deliberations, *the agents who select the principles that are to govern the basic structure of their* social life (DL 572; emphasis added).

Moreover, in this later discussion, Rawls frequently characterizes the parties in the original position in terms similar or identical to those which characterize the members of the well-ordered society (see, for example, the description of each as self-originating sources of valid claims at DL 548, 564, and 543 respectively).

From all these claims jointly, we are quickly led to the conception of particular individuals, mature and partially formed by their own pasts and the previous conditions of their society, who voluntarily come together and, temporarily assuming the constraints and veil of ignorance of the original position, choose principles that are henceforth to govern their claims upon one another. The veil of ignorance is then lifted gradually, in accordance with the four-stage sequence. These individuals recover knowledge of themselves, their pasts, their habits, interests, and conceptions of the good, and immediately proceed to realize their chosen well-ordered society in conformity with the two principles of justice. This is the conception the continuity thesis expresses.

The truth of the continuity thesis gives credence to an implication common to each of the early criticisms of Rawls mentioned earlier. Schwartz and Nagel both claimed that someone of a strongly socialist or communitarian persuasion might be frustrated in her efforts to realize her conception of the good in Rawls's well-ordered society. Gauthier argued that individuals with the economic rationality Rawls ascribes to the parties in the original position would repudiate the two principles of justice for others that better enabled them to pursue their individual interests. And similarly, Miller's criticism can be understood as suggesting that someone with a highly individualistic, ruling class-determined conception of the good might be frustrated in realizing it by choosing the difference principle. Now if the continuity thesis is true, the parties in the original position must at least consider the possibility that in fact they may be any of these types of people, in addition to numerous other possibilities (for example, that their conception of the good includes seeing other persons in positions of the lesser liberty). They must consider the possibility that by choosing as they do – *however they choose* – they risk at the very least the extreme frustration of their deep-seated desires and conceptions of the good; or at most their gradual extinction as continuing personalities

[21] Rawls, *Dewey Lectures*, *ibid*. Page references will be in the text, preceded by DL.

with identifiable tastes, interests, and values. Thus they must be prepared, as we have seen a transpersonally rational moral point of view requires, to give up all that is central to their prior sense of self for the sake of those principles which they agree are to regulate their behavior. But since they are also assumed by Rawls to be primarily concerned in the original position to advance their own interests and conceptions of the good in the subsequent society, it is difficult to see how both conditions can be satisfied; and how they can therefore choose principles of justice under the constraints of the original position at all. Rawls's Humean commitment to principles of egocentric rationality comes into direct conflict with the transpersonally rational moral point of view he so eloquently describes at TJ 537.

7. Rawls's Instrumentalism

But now suppose this dilemma solved, as do Rawls's critics. Suppose, that is, that the truth of the continuity thesis is consistent with the parties' choice of *some* principles of justice that govern their society once the veil of ignorance is lifted. In this case, there are implications both for Rawls's essentially Instrumentalist justification of the two principles of justice, and for the method of wide reflective equilibrium in which they are embedded. In *A Theory of Justice*, Rawls seems clearly committed to justifying these principles as necessary, instrumentally rational means for furthering the interests of free and instrumentally rationally persons (TJ 11, 16, 47, 94, 172), regardless of what these interests are (TJ 129; also see 432). In determining the basic structure of the well-ordered society, these principles indirectly constrain the interests, aspirations and conceptions of the good of its citizens, and so the actions they take to realize them:

> In justice as fairness, persons *accept in advance* a principle of equal liberty ... They implicitly agree, therefore, to conform their conceptions of their good to what the principles of justice require, or at least not to press claims which directly violate them (TJ 31; emphasis added).

Rawls's concept of the original position is thus designed to produce the choice of the two principles of justice as an outcome of the parties' recognition in the original position that under conditions of moderate scarcity of resources to which all have a prima facie equal claim, this is the most instrumentally rational way for each of them to secure their own interests *in the resultant well-ordered society* (TJ 119; also see Section 22).

Now the continuity thesis also implies that from the point of view of any single individual in the original position, the two principles of justice are necessary means to the realization of his system of ends under the circumstances of justice only if, when the veil of ignorance is lifted, and he discovers his own conception of the good, he has no cause to regret his choice of the two principles, no matter what that conception of the good turns out to be (TJ 421-22). But as we have already seen, Rawls himself acknowledges that

this may not be the case (TJ 176). Moreover, the beneficiaries of unjust institutions also know that their inability to maintain their favored positions is a direct consequence of having chosen principles of justice which, when implemented, so alter the circumstances of their life that their original plans of life, i.e. continuing to benefit from these institutions, become untenable. But we have also seen that if the parties know, in the original position, that implementation of the principles of justice *may* thus require them to sacrifice rather than advance their conceptions of the good in this way, they cannot fail to see the instrumental irrationality of choosing the two principles of justice in the first place.

It may seem that the parties do not have reason to regret their choice if they recognize that the two principles of justice were the best available alternative open to them.[22] And perhaps it is true that no alternative principles of *justice* would have the effect of securing their future happiness and security. But it was open to them not to choose principles of justice at all, i.e. to opt for some version of the "No Agreement Point" (TJ 147). Unless we assume that the parties were forced into the original position – surely an unpalatable assumption in view of the parties' freedom and autonomy (TJ 11, 13), and Rawls's allegiance to traditional Social Contract Theory, it is open to the parties to regret choosing to live by principles of justice in the first place, rather than to maintain the status quo in their previous society.

Rawls does not consider the latter as a viable alternative for the parties in the original position, but there is no clear reason why he should not. For unlike traditional Social Contract Theory, the parties do not, on the continuity thesis, enter into the original position from a state of nature mutually acknowledged as unacceptable. So it is consistent with the constraints on information expressed by the veil of ignorance (TJ 12, 136-7), i.e. that the parties know nothing of the circumstances of their own society, that the parties nevertheless elect to take their chances in their society as it has been up to now, rather than risk having to abdicate everything that gives meaning and satisfaction to their lives – even if this requires the deliberate perpetuation of social injustice.

For there is, in addition, nothing in the description of the parties' motivation, circumstances, or interests in *A Theory of Justice* that commits them to choosing *just* principles for society. Recall that their stipulated sense of justice required only that they be able to recognize and honor principles of justice once chosen, not that they deliberately set out to choose justice in the first place. By hypothesis, the parties are moved by the desire to further their conceptions of their own interests, or good, *whatever* this may turn out to be (TJ 129). And the fact that it is the circumstances of justice that move them to deliberate (TJ 128) does not imply that they must opt for just principles to

[22] I was helped by discussion of this point with John Rawls.

adjudicate their claims. Rawls often talks as though the specification of the circumstances in which questions of justice arise naturally generate a motivation to seek a just resolution of the conflict of interests that characterize this situation (TJ esp. 126-7). But there is no reason to assume this. An instrumentally rational individual whose interests are already immoral or unjust has no *prima facie* reason even to view those circumstances as circumstances of justice, for there may be more efficient ways of removing any obstacles to the achievement of her ends. In the face of such considerations, one's sense of justice might well remain dormant.

Thus for a continuing, instrumentally rational individual whose plans and projects already violate the moral constraints embodied in the two principles of justice, these principles are not necessary means but rather obstacles to the achievement of those plans. Since all parties in the original position must canvass in advance the possibility of being precisely such individuals when the veil of ignorance is lifted, they must view these principles as expendable, just as are any other inefficient means to the instrumentally rational achievement of one's ends; and so reject them accordingly.

The general problem begins to emerge. In order to motivate uniquely the choice of principles of *justice*, Rawls must presuppose that the parties in the original position have not only a conception of the good they want to advance and a *sense* of justice, but in addition a motivationally effective *desire for or interest in* justice. Otherwise the two principles of justice are not instrumentally rational for them to choose.

Now in the *Dewey Lectures*, Rawls reformulates the parties' motivation in the original position in order to meet this requirement. There the parties are described as being moved by their highest-order interest in developing and exercising their sense of justice (DL 525-6). First, this means that whatever else they want, they know at least that they want to be moral persons. This reformulation thus purchases motivation to choose principles of justice at the cost of attenuating the opacity of the veil of ignorance. More seriously, it attenuates the ability of the original position to provide an independent, choice-theoretic justification[23] for the powerful moral conception the well-ordered society expresses. As we have seen in Chapter IX.4.2-4, it purchases moral unanimity at the cost of objective validity.

To see this, reconsider the justificatory role of the circumstances of justice in determining the choice of the two principles of justice in *A Theory of Justice*. Here the idea is that the original position represents the salient features of the

[23]Rawls states his commitment to this kind of justification in many places. See TJ 16, 47, 94, 119-121, 125,172, 583. He retracts it at DL 572, and again more forcefully in "Justice as Fairness: Political not Metaphysical," 237, fn. 20 (*op. cit.* Note 19), and yet again in *Political Liberalism* (*op. cit.* Note 3), 53 fn. 7.

situation in which equally positioned agents compete for scarce resources to further their unstated, mutually conflicting ends. These circumstances are compelling to us because most of us understand what it means to have to compete for scarce resources in our daily lives. To show that these circumstances plus the other special conditions that define the original position generate certain principles of justice would provide a powerful incentive for us to accept them, for it would depict those principles as a rational outcome of conditions essentially reflected in our experience. It would thereby appeal to our actual life situations, independent of the particular values and moral convictions any of us may happen to have; and in so doing, gain objective authority for each of us. This is what an objective moral justification should do.

By contrast, the reformulations in the *Dewey Lectures* require that each of the parties' conceptions of the good be constrained by their shared highest-order interests in developing and exercising their senses of justice. To stipulate at the outset that the parties are *overridingly* motivated to act on principles of justice is to ascribe to them a motivation that *a fortiori* overrides the motivational incentive of the circumstances of justice. This means that each of the parties is more concerned to develop and exercise her sense of justice than she is to acquire sufficient resources to further the other ends that are peculiarly hers. That is, each party wants to become just more than anything else, and everything else she wants is subordinate to this one. Second, the parties must therefore share the interest of developing and exercising their senses of justice as a *nonconflicting* highest-order interest. Third, this interest must be partially *determinate*, since they must therefore have some prior conception of what justice is, in order to be moved to develop and exercise their sense of it (to see this, try substituting "sight" for "justice").

But the relative determinateness of the parties' conception of justice in turn further undermines the sense in which they can be said to *choose* just principles through rational negotiation, consistently with the assumption of pure procedural justice (TJ 120, 136). It is hard to see how a shared antecedent desire to develop and exercise one's sense of justice, of which each party has an antecedent conception, could fail to influence the outcome of the procedure of rational deliberation that the parties undergo in the original position; and how that antecedent conception of justice could fail to bias the process of deliberation as well. These revisions thus move Rawls's account closer to a conception of perfect procedural justice, in which we begin with an antecedent conception of what justice requires and rig the procedure so as to produce that outcome.

These revisions also undermine Rawls's claim, first made in *A Theory of Justice* (128) and developed more fully in the *Dewey Lectures* (557-60, 561, 564, 571), that the parties are not bound by antecedent moral ties. How can they have an antecedent conception of justice and a shared, overriding interest in

developing and exercising their sense of it without being bound by antecedent moral ties? This overriding, nonconflicting interest in justice *is* an antecedent moral tie. To each party, it stipulates not that that individual should try to obtain as large a quantity of primary goods as he can, but rather that that same individual should see to it that each and every one of the parties, including himself, should obtain a fair allocation of primary goods. That is, it requires each party in the original position to be cognizant of the well-being of all of them, thereby violating the stipulation of mutual disinterest. Furthermore, the more fully the sense of justice is conceived as a motivationally effective intentional object for the parties in the original position, the more motivationally otiose the subjective and objective circumstances of justice become – and the less the rational appeal of Rawls's argument to the unconverted, the cynical, the opportunistic, or the unabashedly self-seeking, all of whom may hold, rather like the Hobbesian Free Rider, the belief that giving primacy to justice for all over the exigencies of personal self-aggrandizement is a naiveté of those with the *noblesse oblige* to indulge it.

Indeed, that the parties are overridingly motivated to realize and exercise the "capacity ... to understand, to apply and to act from (and not merely in accordance with) the principles of justice" (DL 525) *tautologically requires* them to choose the two principles of justice in order to realize this aim. Rawls has ensured that it is instrumentally rational for the parties in the original position to choose the two principles of justice by stipulating in advance that this is what they are most highly motivated to choose, regardless of the further ends they serve. He has built the choice of the two principles of justice into the original position in such a way as to effectively nullify the force of their instrumental rationality altogether.

Thus Rawls has in effect dropped the Instrumentalist strategy of justification and substituted a more purely deductive one, just as he originally intended (TJ 119-120). The two principles of justice are, in the *Dewey Lectures*, no longer justified as instrumentally rational means to the promotion of even a circumscribed range of conceptions of the good. They are stipulated to be final ends, i.e. part of each of the parties' conceptions of the good itself, in the premises from which they are then deduced as outcomes of deliberation.

This stricter form of Deductivism creates its own problems. Whereas Gewirth began with very weak premises and had to add stronger, more value-laden assumptions to the argument as it proceeded in order to derive his principle of generic consistency as a conclusion, Rawls adds stronger, more value-laden assumptions to his premises – the conception of the original position – in order to preserve the value-neutrality of the derivation itself. But this makes it harder to conceive the original position itself as generating, rather than presupposing, the moral ties it was its original function to justify.

So either the parties must be viewed as motivationally uncommitted to making a specifically moral choice,[24] in which case a return to the unjust status quo is a viable alternative; or else we must assume *a priori* that they are morally motivated to choose the two principles of justice, in which case it remains an open question how those principles are to be independently and objectively justified. As we have seen in Chapter IX.4, this choice of alternatives is inevitable, and fatal, for the committed Instrumentalist.

But how committed an Instrumentalist has Rawls ever been? Is there any remaining justificatory force in the original position for the two principles of justice, now that Rawls has made this recent modification? The answer is yes. However, this force does not derive from the demonstrated instrumental rationality of the two principles of justice; for as we have seen in Chapter IX.4, it is vacuously true that an agent will choose what she has special motivation to choose, other things equal. This fact cannot justify that choice to us unless we, too, have that special motivation, or would have it under appropriate circumstances.

Now we have just seen that the relevant circumstances on which the significance and justificatory force of the original position originally relied were supposed to be the more practically compelling *circumstances* of justice, in which the parties were realistically portrayed as competing on an equal basis for scarce resources for achieving their ends, not a highest-order interest in their sense of justice. And it is controversial whether this moral motivation is as essentially reflected in our actual lives; i.e. whether it is in fact more important to us to obtain for ourselves an equitable allocation of the scarce resources we need to survive comfortably, or to develop and exercise our sense of justice. This moral motivation may be more an expression of our idealized self-conceptions than well-grounded in a conception of the self that is adequate to the psychological facts about us.

If the two principles of justice are to be generated by a highest-order interest that we do not actually have, or have only to a relatively minor degree, then as we have seen in Chapter IX.4.3, the extent of their persuasive force for us will be inversely proportional to the extent of the remoteness of the ideal self-conception that includes that interest from our actual emotions and dispositions. In this case, the justification of the two principles will require that we be convinced, first, that the parties in the original position represent characteristics and dispositions that we have, or do or should aspire

[24]This possibility is supported by the stipulations that (i) the parties are mutually disinterested (TJ 13, 127); and (ii) they are not bound by prior moral ties (TJ 128). If these things are true of them, then why should they commit themselves to choosing principles of justice when each could stand to benefit from injustice? Clearly the additional stipulation of risk aversion is inadequate to answer this question, since each might value the benefits of injustice highly *enough* to risk being victimized by it.

to have; and second, that the well-ordered society, structured by the two principles, depicts the kind of society we envision as a felt social ideal.

That is, both the original position and the well-ordered society now represent deductively connected but value-laden hypotheses that must be evaluated for their rational persuasiveness from a third perspective, namely that of the engaged but as yet uncommitted reader for whom the question in what a just society consists is a pressing one. Thus the effect of Rawls's recent revisions in his conception of the original position is to shift the fulcrum of his justificatory strategy from the choice of the parties in the original position to the receptiveness of the reader to the values the original position and well-ordered society embody. Let us call this the *rational reader conception of justification*.

Rawls has developed the rational reader conception of justification in recent writings, by explicitly addressing as his audience those who affirm the liberal-democratic tradition his theory of justice also affirms. Clearly, he means to be addressing us as readers about issues of central importance to us, and articulating the basic conditions under which public agreement among us can be achieved:

> [T]his conception [of justice] provides a publicly recognized point of view from which all citizens can examine before one another whether or not their political and social institutions are just. It enables them to do this by citing what are recognized among them as valid and sufficient reasons singled out by that conception itself.... *[J]ustification is not regarded simply as valid argument from listed premises, even should these premises be true. Rather, justification is addressed to others who disagree with us, and therefore it must always proceed from some consensus, that is, from premises that we and others publicly recognize as true;* or better, publicly recognize as acceptable to us for the purposes of establishing a working agreement on the fundamental questions of political justice (italics added).[25]... [Kantian

[25]"Justice as Fairness: Political not Metaphysical," 229. I do not agree with Rawls that public recognition of such premises "as acceptable to us for the purposes of establishing a working agreement on the fundamental questions of political justice" is better than premises "that we and others publicly recognize as true." Everyone recognizes some premises as true. The question is whether these premises – the ones to which we are most deeply committed psychologically as well as epistemically, and which Rawls describes as "comprehensive doctrines" in *Political Liberalism* (*op. cit.* Note 3) – are to be examined, criticized and revised in light of public discussion of them, or excluded – perhaps shielded would be a better word – from such discussion altogether. By leaving unmentioned, unscrutinized, unanalyzed, unevaluated, and unrevised what one in fact actually does recognize as true, the former, merely pragmatic public recognition precludes authentic Socratic dialogue and the genuine meeting of minds that can result from it. These values are discussed in Chapter I, Section 2, above. In *Political Liberalism* Rawls initially advocates a "don't ask, don't tell" approach to our deepest moral and epistemic convictions (PL 15-16, 152, 153); but later distinguishes between exclusive

> constructivism] recasts ideas from the tradition of the social contract to achieve a practicable conception of objectivity and justification founded on public agreement in judgment on due reflection. The aim is free agreement, reconciliation through public reason.[26]

Since the two principles of justice are no longer justified as instrumentally rational for the realization of the parties' individual conceptions of the good in the well-ordered society, this more recent conception of justification does not require the truth of the continuity thesis. So we must now turn to the question of whether there are sufficient resources in Rawls's theory of justice to replace it.

8. The Discontinuity Thesis

There is some textual evidence in *A Theory of Justice* that undermines the continuity thesis. The falsity of the continuity thesis is, for example, suggested by Rawls's emphatically drawn distinction between the parties in the original position as "theoretically-defined individuals" and the actual propensities of people in everyday life (TJ 147); he asserts that the mutual disinterest in one another of the parties is not necessarily continuous with the motives of "persons in everyday life who accept the principles that would be chosen and who have the corresponding sense of justice" (TJ 148). One might speculate that by "everyday life" Rawls may mean here not only the everyday life of the reader, but perhaps also the everyday life of a person in the well-ordered society. In any case, the latter point is made more strongly in a later discussion of this issue:

> [T]he motivation of persons in a well-ordered society is not determined directly by the motives of the parties in the original position. These motives affect those of persons in a well-ordered society only indirectly: that is, via their effects on the choice of principles. It is these principles, together with the laws of psychology (as these work under the conditions of just institutions), that determine the resulting motivation.[27]

These passages taken conjointly do not, of course, explicitly conflict with all the clauses of the continuity thesis. They are primarily aimed at confuting the purported continuity of the original position's individualistic motivational assumptions with the underlying psychology of members of the well-ordered society. But in so doing the point is also made that the parties in the original

---

public reason as appropriate to the ideal case (PL 248), and inclusive public reason which introduces discussion of comprehensive doctrines in non-ideal cases of serious political dispute or destabilizing conflicts about constitutional essentials; and later still, in the Introduction to the Paperback Edition, admits the introduction of comprehensive doctrines into public discussion at any time (li-lii).

[26] *Ibid*., 230.

[27] "Fairness to Goodness," *The Philosophical Review 84* (October 1975), 543.

position are psychologically discontinuous with members of the well-ordered society. So whatever prior conceptions of the good, tastes, and interests the parties may have held, their subsequent psychological proclivities nevertheless conform to the constraints of the two principles of justice. Thus we can read these passages as conflicting with clauses (2) and (3) of the continuity thesis.

Clause (1) is called into question by Rawls's claim that "We use the characterization of the persons in the original position to single out the *kinds* of beings to whom the principles chosen apply" (TJ 505; emphasis added). If we understand this to mean that the parties in the original position are not among those *particular* individuals to whom the principles of justice might apply, but rather schematic adumbrations of the *type* of individual for whom they are intended, i.e. moral persons, there is no reason to suppose that the parties bear anything like the concrete relation to particular members of the well-ordered society suggested by the passages adduced in Section 6. On the present construal, the parties in the original position are as much hypothetical constructs to the members of the well-ordered society as they are to the reader; and indeed there is some further evidence to support this possibility.

Near the beginning of *A Theory of Justice*, Rawls relates his conception of the original position to the hypothetical state of nature characteristic of traditional Social Contract Theory. After briefly limning what will later become the four-stage sequence in which a conception of justice, then a constitution, and a legislature to enact laws is chosen, Rawls then argues that

> *Our* social situation is just if it is such that by this sequence of *hypothetical agreements* we would have contracted into the general system of rules which defines it. Moreover, ... it will then be true that *whenever* social institutions satisfy these principles, those engaged in them can say to one another that they are cooperating on terms to which they *would* agree if they were free and equal persons whose relations with respect to one acceptance of the corresponding principles of justice. *No* society can, of course, be a scheme of cooperation which men enter voluntarily in a literal sense; .... Yet a society satisfying the principles of justice as fairness *comes as close as a society can* to being a voluntary scheme, for it meets the principles which free and equal persons *would* assent to under circumstances that are fair (TJ 13; emphasis added).

The hypothetical nature of the original position with respect to the vantage point of the reader is reaffirmed in a number of places in *A Theory of Justice* (16, 21, 48, 115, 120, 167, 587). The importance of the above passage lies instead in the facts that first, the continuity thesis is vigorously rejected in its entirety; and second, the implied relation between the parties in the original position and the members of the well-ordered society is very different here from what it is elsewhere. Here the implication is not that the parties afterwards "return to their place in society and henceforth judge their claims

on the social system by these principles" (TJ 196) with the newly requisite attitudes. Rather, it is suggested that the original position is equally hypothetical relative to any society, including the well-ordered one; and that the former functions as an idealized standard in comparison with which the underlying principles, constitution, legislature, and laws of such a society can be appraised, criticized, and, in theory, revised.

On this construal, we regard the well-ordered society not as an immediate and concrete temporal consequence of the series of decisions made throughout the four-stage sequence of the original position and then miraculously implemented by the subsequent efforts of the parties. On the contrary, we view the well-ordered society as the theoretically projected outcome (remote though it may be) of our concerted application of a *device*, i.e. the hypothetical original position and the principles chosen there, to the circumstances of moral and political conflict that regularly confront us. This device both insures the impartiality of our moral judgments and also yields substantive moral principles in accordance with which we can judge these issues.[28]

Similarly, in the *Dewey Lectures*, Rawls frequently refers to the original position as a construction (DL fn. 41, 532), a point of view (DL 554, 560, 567), and a framework of deliberation (DL 533, 560, 561); and the parties themselves as "agents of construction" (DL 547, 552, 560). These characterizations reinforce the interpretation of the original position as a theoretical construct that enables us as well as citizens in the well-ordered society to perform systematic moral deliberation. And on page 234 of "Justice as Fairness: Political not Metaphysical," Rawls characterizes his conception of the parties in the original position as "a basic intuitive idea assumed to be implicit in the public culture of a democratic society;" this is part of the "publicly recognized point of view" that enables us to evaluate the justice of our own institutions. He also, at last, describes it emphatically as a "device of representation" (236-8) within which

> the conception of justice the parties would adopt identifies the conception we regard – *here and now* – as fair and supported by the best reasons.... As a device of representation the idea of the original position serves as a means of public reflection and self-clarification.... The original position serves as a unifying idea by which our considered convictions at all levels of generality are brought to bear on one another so as to achieve greater mutual agreement and self-understanding (238; emphasis in text).

In these passages the detachment of the parties in the original position as a device we use from the citizens of the ideal well-ordered society generated by

[28]Something like this strategy seems to lie behind Rawls's treatment of substantive questions in the second part of *A Theory of Justice*, e.g. civil disobedience (Secs. 55-59), but this is too large a topic to discuss here.

it is almost complete: Each bears a continuity relationship, not to the other, but to us.[29]

Thus these passages suggest an alternative to the continuity thesis in the form of the following conditions:

> (1') The parties in the original position do *not* regard themselves as future members of the well-ordered society (clauses (1) and (3) of the continuity thesis). Rather, they recognize themselves to be psychologically discontinuous with those members, and recognize also that their choice of principles determine the general motivational features of the members of the well-ordered society.
>
> (2') The original position itself is *not* an actual event (clause (2) of the continuity thesis), but rather a hypothetical one relative to the well-ordered society. Hence the parties are not in fact physically continuous with any member of the well-ordered society.

Let us call this the *discontinuity thesis*.

A number of exegetical consequences follow from adopting the discontinuity thesis as the favored interpretation of the original position. If we assume the falsity or irrelevance of the continuity thesis to the Rawlsian enterprise, those criticisms that presuppose it must be disregarded. One cannot then argue against the conception of the original position or the two principles of justice chosen there on the grounds of what the parties so situated would do or think after they got out of it or before they went into it,[30] or how their society might be colored by their prior psychological proclivities. Without the continuity thesis, there is no *prima facie* reason for the parties in the original position to suppose that anyone's needs or conception of the good might conflict with, undermine, or be frustrated by the constraints imposed by the two principles of justice on the basic structure of society.

This is not to propose that everyone's psychology in the well-ordered society must be consonant with these principles. It is just to argue that without the notion of continuing persons, this possibility does not suffice to deter the parties in the original position from choosing as they are presumed to do. It does not suffice because the parties then have no reason in deliberating to provide for the possibility that the persons they turn out to be might be persons to whom their choice in the original position is unacceptable. If the continuity thesis fails, the parties in the original position

---

[29] Cf. Footnote 19.

[30] Certain of Ronald Dworkin's arguments in "The Original Position" (*University of Chicago Law Review 40*, 3 (Spring 1973), 500-33; reprinted in Daniels, 16-52) against the justificatory function of the original position that depend on his distinction between "antecedent" and "actual" interest (20-21) would necessitate revision on this interpretation.

must be regarded as self-determining in the very strong and liberal sense that in the original position, they determine the kinds of persons they are to be, the kind of psychology they will have, and the kinds of moral constraints they will be prepared to accept, by deciding what principles of justice are to regulate their interactions. The circumstances of the original position must then be viewed as a radical discontinuity in their adult lives, after which they become the kinds of persons who are constrained and partially determined, not by the continuity of their previous psychological histories, but by their choice of moral principles in the original position.

At the same time, those passages we have cited from *A Theory of Justice* and *The Dewey Lectures* that lend support to the continuity thesis must be similarly bracketed. We must, for example, interpret Rawls's discussions of the expectations of the parties in the original position in the same hypothetical light as we do the concept of the original position itself. The parties must be conceived, and must conceive themselves, as deciding on principles *as though* such principles were to govern their life prospects. For they recognize that they are in fact choosing principles not for themselves, properly speaking, but for the persons they thereby choose to become. Thus they must regard themselves as advancing in the subsequent society not only their conceptions of the good, but indeed their idealized self-conceptions which the choice of principles determine.

Rawls's arguments regarding the strains of commitment must be qualified in much the same way: The issue then becomes not whether the parties can adhere to the chosen principles, but instead whether the preferred self-conception includes this capacity. This implies, first, that the capacity for a sense of justice cannot be stipulated as a motivational assumption of the original position, independently of this preferred self-conception. Secondly, it implies that the capacity for a sense of justice cannot be used as a criterion for differentiating between acceptable and unacceptable principles of justice. For we can expect a great variety of such principles to be successful in tailoring a self-conception that will stably adhere to them.

Finally, the abandonment of the continuity thesis entails the abandonment of the Instrumentalist strategy of justification that is, for many, centrally definitive of the social Contract-Theoretic tradition. That tradition is founded on the reasoning that a justified society is one by the rules of which individuals who are instrumentally rational and self-interestedly motivated to improve their lot in the state of nature would agree to be bound in it, in order to regulate their interactions. In this picture, the state of nature, the self-interested and instrumentally rational individuals, and the social contract are jointly continuous but hypothetical relative to our actual society. But by abandoning the continuity thesis, the modifications we have traced in Rawls's later writings eliminate the state of nature, the self-interest, and the instrumental rationality of the individuals; and stipulate the hypothetical

nature of the contractual agreement in the original position relative to the ideal well-ordered society itself. There are many elements in Rawls's normative theory of the well-ordered society that continue to affiliate him with the tradition of Social Contract Theory. But his developed conception of its metaethical justification represents a radical departure from that tradition.

9. Personal Identity and Wide Reflective Equilibrium

Next I consider the implications of the discontinuity thesis, first, for Rawls's views on personal identity; and second, for his concept of wide reflective equilibrium. Samuel Scheffler's criticism of Rawls's theory has focused on the seeming disparity between claims made in *A Theory of Justice* supporting the choice of the two principles of justice over Utilitarianism, and Rawls's more recent treatment of the issue of the relevance of the problem of personal identity to moral theory.[31] In *A Theory of Justice*, it was suggested that the fact that Utilitarianism does not take seriously the distinction between persons might be a reason why the parties in the original position would be disinclined to choose it. For they would then have good reason to doubt whether a society erected on Utilitarian first principles would protect and promote those long-term plans and interests which the parties each know themselves to have (TJ 27-29). In "The Independence of Moral Theory," on the other hand, Rawls is concerned to show that conclusions in the philosophy of mind concerning personal identity, i.e. that it involves bodily continuity and also mental continuity subject to varying degrees of fluctuation, do not conclusively favor one moral theory over any other. Thus Rawls claims that

> what sorts of persons we are is shaped by how we think of ourselves and this in turn is influenced by the social forms we live under.... There is no degree of connectedness that is natural or fixed; the actual continuities and sense of purpose in people's lives is relative to the socially achieved moral conception.[32]

As we have seen, this argument provides important support for his view that the parties in the original position are psychologically discontinuous with the members of the well-ordered society.

But Scheffler has tried to show that this has one of two equally problematic implications. One possibility is that the parties then cannot assume themselves to have long-term plans and purposes – since this feature would characterize a particular kind of personal identity which is no more natural or fixed than a weaker one in which plans, projects, memories, and experiences undergo continual alteration and replacement. Hence they cannot choose principles of justice with an eye to protecting such long-term plans

[31]Samuel Scheffler, "Moral Independence and the Orginal Position," *Philosophical Studies 35*, 4 (May 1979), 397-403.
[32]*0p. cit.* Note 18, p. 20.

and purposes. The second possibility is that the parties' choice of the two principles of justice, on the supposition of having long-term interests, demonstrates only that individuals having a certain kind of personal identity would choose a society that would protect it, without providing any independent argument against the choice of Utilitarianism.[33] If the parties are then *not* assumed to have long-term plans and purposes but nevertheless *are* assumed to choose principles of justice for the basic structure of the society in which they will then live, it is then an open question whether they would choose a society which protected long-term interests over one that did not.

But if the continuity thesis is supplanted by the discontinuity thesis, and in particular clause (1'), these difficulties do not arise. For it is only if the parties are conceived as continuing persons who had adopted certain projects and purposes prior to the original position, which they then advance in the well-ordered society subsequent to it, that there is any independent requirement for how long a person in the original position must endure, and how long a long-term interest must be in order to count as long-term. A person, and hence her goals and interests, must endure long enough to have originated and engendered in the person a deep commitment to the fulfillment of these goals and interests before the circumstances of the original position occurred; they must survive the protracted period of conflict, dialogue and deliberation which the original position, with the support of clause (2), surely entails; and it must survive the actual lengthy period of implementation of the two principles of justice in the well-ordered society which the continuity thesis plus the four-stage sequence entails. Such longstanding commitment to a goal or interest is impressive indeed. It might even survive what we normally think of as a natural human lifespan.

Rejecting the continuity thesis, on the other hand, permits us to leave open the question of how long, in actual time, the parties' long-term interests must be in order to count as long-term. It is then sufficient that a long-term interest survive for the duration of a person's adult life, as we would normally expect. But there is now no reason to place any prior constraints on how long such a life must be. Hence whether a person has a weak identity or a strong one is irrelevant to whether that person can be said to have long-term interests or not. The person's interests are identified as long-term relative to the duration of her personal identity. Now since there are no longer any independent constraints on how long the parties themselves endure, nothing about the conception of the original position forces the characterization of the parties as having either particularly weak or particularly strong personal identities. And the ascription to them of long-term interests fails to decide this question one way or the other.

---

[33]Scheffler, *op. cit.* 399-401.

Furthermore, that a person has a weak personal identity in any case does not imply a lack of concern with personal survival. Even if I know that my character is so volatile and unstable that I can realistically expect to be a completely different person in five years, I need not be happy about this. Or I may wish my interests to endure for as long as I do, however long that is. It is both conceivable and likely that an individual with a weak personal identity would be either concerned with her own personal survival, or hold personal survival as a value in general, or both. And so long as the continuity thesis is rejected, the question of how long, in real time, such an individual would consider it in general valuable for a person to survive, is once again left open. So the fact that the parties in the OP may have weak personal identities does not imply that they would have no reason to choose principles of justice that would respect and protect long-term interests. Indeed we might expect the concern with person survival to increase with the threat to personal survival. Thus that Utilitarianism fails to take seriously the distinction between persons and hence would fail to protect their interests remains a good reason for any party in the original position not to choose it.

Now clause (1') of the discontinuity thesis implies that the parties know they will not psychologically survive past the circumstances of the original position. They regard themselves as determining future selves for the well-ordered society in light of the chosen principles of justice, and hence as determining the long-term goals and interests these selves will have. Hence they choose principles, not with an eye to promoting instrumentally their own long-term interests, but rather with an eye to protecting the long-term interests of the kind of person they simultaneously choose to become. And to suppose that the parties could be unconcerned about the survival of those long-term interests – however long in real time they might be – would be to suppose that they were indifferent to the particular nature of the self they had chosen. But since they regard themselves as self-determining, there is no reason to believe they would be.

Thus Rawls's claim, that conclusions about the nature of personal identity are irrelevant for the construction of a moral theory, can be made to hold in spite of the apparent conflict with the earlier argument against Utilitarianism. These conclusions are irrelevant as long as the characterization of the original position is not complicated by the assumption of the continuity thesis. For only then is Rawls committed to a type of personal identity which the parties might desire self-interestedly to prolong into the well-ordered society, rather than one they desire disinterestedly to create.[34]

[34]However, my conclusions here should not be taken to endorse Rawls's later disavowal of the relevance of metaphysical questions to his theory of justice *tout à fait* ("Justice as Fairness: Political not Metaphysical," 230, 238 40, fn. 22). If my arguments are well taken, Rawls's disavowal was too sweeping.

But perhaps the most important consequence of supplanting the continuity thesis is that greater attention needs to be directed towards Rawls's concept of wide reflective equilibrium. In *A Theory of Justice*, the concept of wide reflective equilibrium referred to a stable state in which the description of the original position and the principles chosen in it to govern the well-ordered society had been mutually adjusted and compared with other alternatives so as to finally match our considered moral judgments (TJ 20, 48-9). As we have seen, many of the features originally ascribed to the original position have required modification in order to circumvent the unacceptable implications of the continuity thesis. In particular, the motivational features of the original position that, at least on the interpretation initially offered in this chapter, lead the parties to choose principles of justice in order to advance their conceptions of the good in the subsequent well-ordered society must be abandoned, and replaced by some other connection between the original position and the well-ordered society. The strictly deductive connection Rawls offers in the *Dewey Lectures* between the parties' motivation and the outcome of their deliberation does not require the continuity thesis for its plausibility.

Thus clause (2') of the discontinuity thesis denies that the major connection between the original position and the well-ordered society is mediated by continuing persons who are assumed to participate in both. And we have already seen that Rawls himself later gave increasing prominence to us, the readers, conceived as members of a liberal democratic society who are reflective, self-critical, and morally concerned thinkers who attempt to give coherence, substance, and reality to their considered moral judgments. As moral mediators between the original position and the well-ordered society, we advert to the original position in order to attain the requisite impartiality of judgment, and to the well-ordered society in order to substantiate and specify the scope of application of those judgments themselves. Hence the attainment of wide reflective equilibrium must be measured by the internal coherence of our own moral judgment on the one hand, and the points of view from which we make them on the other. The importance of this line of thought to Rawls's thinking from the very beginning is strongly suggested by his closing remarks in *A Theory of Justice*.[35]

[35]For this reason I now think I was too hasty in claiming (in "A Distinction Without a Difference," *Midwest Studies in Philosophy VII: Social and Political Philosophy* (Minneapolis, Minn.: University of Minnesota Press, 1982), p. 406) that Rawls's theory of justice contains no action-guiding part at all. Although this was true of *A Theory of Justice*, there has always been a practical and applied strain in his thought that became increasingly salient in his later writings.

10. Moral Objectivity and Pure Procedural Justice

We are now in a better position to answer the questions raised in Section 5.2. There we asked whether Rawls's conception of rational deliberation in the original position in fact qualifies as an instance of pure procedural justice, such that the chosen principles are defined and recognized as just because and only because of the rational procedure by which they were generated, irrespective of prior, prereflective moral intuitions that might conflict with them. We also asked what, in the event of such conflict, would need to be revised: the conception of the original position? The Instrumentalist deliberation procedure? The two principles of justice and the well-ordered society they structure? Or the foundational moral intuitions that anchor the entire metaethical scheme that Rawls offers? Finally, we noted that if these anchoring intuitions remain the final arbiter against which all other elements of Rawls's theory must be measured, then Rawls's deliberative procedure is not one of pure procedural justice after all. For these intuitions, however inchoate, in effect constitute a prior conception of justice that, when elaborated through the process of wide reflective equilibrium, constrain and determine the procedure we construct so as to arrive at it.

We saw in Section 5.1 that Rawls originally introduced the notion of pure procedural justice in order to capture an insight about the nature and status of the Instrumentalist deliberations by which the parties in the original position arrived at the two principles of justice in order to advance their individual conceptions of the good. The rational procedure constitutive and definitive of just principles for the distribution of primary goods was the procedure of instrumentally rational choice the parties undertook. We have seen in Sections 6 and 7 that this procedure, under the conditions Rawls originally specified, generates a contradiction in the motivational conception of the original position, and so must be abandoned.

However, we have also seen in Sections 8 and 9 that Rawls in later writings reorients his focus on the metaethical justification of the two principles. He replaces the centrality of the procedure of instrumentally rational choice with a new emphasis on the more purely deductive procedure of theoretically rational theory-construction expressed in the process of wide reflective equilibrium. In this latter process, we as rational readers take a more active role in a more collaborative process of arriving at principles of justice to govern our society: We make adjustments and revisions in all three elements – the initial premises expressed in the original position, the statement of the two principles express in the conception of a well-ordered society, and the commonsense moral intuitions by which we guide our reflections – so as to maximize coherence and consistency among all three. This is exactly what commentators on Rawls's *Theory of Justice* have done, and we have seen that he has taken their criticisms very much to heart. We saw in Section 5.2 that

this process is in fact much closer to the rational scientific procedure analogy with which supplied Rawls his original inspiration in 1951.

How should we then evaluate each of these two procedures relative to Rawls's identification of his view as an instance of pure procedural justice? Since the instrumentally rational procedure of deliberation undertaken by the parties in the original position leads to a contradiction, it fails to meet Rawls's characterization of pure procedural justice, by definition. Rawls's de-emphasis and indeed rejection of this procedure in later writings, and subsequent elaboration of our participation in delineating the conception of a well-ordered society which forms one of his last projects in *Political Liberalism*[36], constitutes a sacrifice of this procedure to the more purely deductive procedure of formulating the premises of his arguments in such a way as to derive analytically what is contained within them, such that both premises and conclusion cohere with and preserve our most central commonsense moral intuitions. This is part of the process of achieving wide reflective equilibrium. Hence just as the analogous procedure in scientific theorizing – in which we, on the basis of our intuitions, formulate the hypothesis and its experimental predictions in such a way as to maximize the likelihood of the latter confirming the former – would count as a case of perfect procedural truth, similarly Rawls's procedure of wide reflective equilibrium would seem to count as one of perfect procedural justice.

However, there is more to the process of achieving wide reflective equilibrium than this. Rawls clearly acknowledges that "a person's sense of justice may or may not undergo a radical shift" as the result of undergoing the process of achieving wide reflective equilibrium (TJ 49). He thereby leaves open the possibility that the commonsense moral intuitions that anchor the process of achieving wide reflective equilibrium may be sacrificed not only to the theory-constructive requirements of coherence and deductive consistency, but also to *the influence of competing moral views with which those intuitions are compared.* So the scientific analogy is rather with the attempt to square one's hypothesis, its experimental predictions, and the underlying intuitions that anchor both with recognizedly anomalous data – perhaps unexpected experimental results, or competing hypotheses that are more powerful or comprehensive – that call all three into question. This process would be one of pure procedural truth. Similarly, the process of achieving wide reflective equilibrium makes our commonsense moral intuitions vulnerable to revision in light of demonstrations that a competing moral theory is, for example, more adept in the casuistry of particularly problematic cases, or more comprehensive in its application, or better grounded in the psychological facts about human beings.

---

[36]*Op. cit.* Note 3.

Hence Rawls's more recent, more purely deductive procedure of justification is an instance of pure procedural justice; but not for the reasons he claims in *A Theory of Justice*. What makes it a case of pure procedural justice is not that the outcome of the parties' deliberation in the original position is by definition just; for we have seen that the conditions under which they deliberate have been revised so as to produce precisely this outcome. What makes Rawls's revised procedure of justification one of pure procedural justice is that *the outcome of the process of achieving wide reflective equilibrium is, according to Rawls, by definition just.* By exposing the process of theory-construction inherent in the process of wide reflective equilibrium to the participation of rational readers who are citizens of a liberal democratic society, we thereby expose it to the very wide range of competing moral views that such a society permits. And by de-throning any individual's commonsense moral intuitions as final arbiters of what such a theory should contain, we signal our willingness to sacrifice those intuitions to the requirements of consistent, reflective, and self-critical rational reflection. This is precisely what the enterprise of Socratic metaethics demands.

We can now see that Rawls's revision of traditional Social Contract Theory is even more radical than it first appeared. We have seen that Rawls supplied Social Contract Theory with a formalizable metaethical underpinning (although not the one with which he began), addressed the question of economic justice that no social Contract Theorist before him had tackled, reinvigorated normative moral theory as a legitimate philosophical endeavor, reincorporated moral psychology as a central part of the Social Contract-Theoretic enterprise, and revitalized discussion of casuistical issues such as civil disobedience, passive resistance, and racism. In addition to this, Rawls ultimately reverses the relation between rationality and power established by Hobbes and rejected by Kant. Whereas traditional Social Contract Theory subordinates unlimited power for everyone in the state of nature to the benefits of instrumental rationality under the social contract, Rawls subordinates the benefits of instrumental rationality under the social contract to the requirements of transpersonally rational procedures of deliberation for establishing its objective validity. He thus answers Nietzsche's devaluation of the character dispositions of rationality with a demonstration that the exercise of power, even under the guidance of instrumental rationality, does not suffice to meet our demand for its objective validity. Only social arrangements justified as the outcome of a collaborative and transpersonally rational procedure of deliberation can do that.

Chapter XI. Brandt's Instrumentalism

Like Rawls, and following in the footsteps of his metaethical strategy, Richard Brandt, too, is an Instrumentalist. His reliance on Instrumentalism undermines his attempt to morally justify his normative theory in a similar manner, even though his normative moral theory is a species of Utilitarianism rather than Social Contract Theory. In this chapter I argue that close examination of Brandt's argument in his *A Theory of the Good and the Right* suggests that the primary determinant of an agent's choice of the Ideal Code Utilitarian Society is not her having undergone cognitive psychotherapy as Brandt claims, but rather her being independently benevolently motivated. But in this case, the concept of instrumentally rational choice is doing no justificatory work. For it is – again – vacuously true that an agent will choose what she has special motivation to choose, other things equal. This fact does not succeed in justifying her choice for us unless we, too, have that special motivation. But if we do, then the argument does not succeed in justifying this choice as instrumentally efficient *whatever* ends we have, i.e. objectively. Brandt's characterization of rational desires as those which survive cognitive psychotherapy suggests a different and more powerful method of moral justification that renders dispensable his Instrumentalist strategy.

Section 1 contrasts Brandt's Instrumentalism with Rawls's. Although they are united in their *de facto* commitment to the Humean conception of the self, Rawls proves to be the more consistent Humean, whereas Brandt incorporates central Kantian tenets, explored further in subsequent sections. Section 2 describes the dilemma generated by his Instrumentalist mode of justification, and also the problems raised by his appeal to the reader's self-interest. Section 3 examines Brandt's analysis of desire, and the Kantian implications of his stipulation that a belief is the precipitating cause of action. Section 4 traces the conditions Brandt imposes on specifically rational desire, with particular attention to his conception of cognitive psychotherapy. Although these retain the universality of his criterion of rational desire, they imply that there are no universally rational desires themselves. Section 5 contrasts Brandt's account of prudence with Nagel's. But we see that Brandt's analysis of prudential motivation in fact accords with Nagel's Kantian one rather than opposing it. Section 6 looks at the irrational desires which Brandt claims cognitive psychotherapy extinguishes, and argues that this criterion does not succeed in distinguishing rational from irrational desires. Section 7 argues that on Brandt's account of benevolence, it is rational neither for a benevolent agent nor for a self-interested one to choose the Ideal Code Utilitarian society. Section 8 concludes that his conception of cognitive psychotherapy has not been exploited to maximum effect.

1. Brandt and Rawls

Ideal Code Utilitarianism describes an ideally rational moral code – a set of moral rules that are "presumably some not very distant variant of some present rules" (290) [1] – conformity to which would maximize welfare for all members of a society. It thus has a long Anglo-American pedigree that extends back to Bentham and Mill. Like Rawls, Brandt tries to justify his normative moral theory to us by arguing that the actions, life-plan, or social arrangements it prescribes can be shown to be the best means to the achievement of an agent's final ends, whatever these may be.

However, whereas Rawls tried to defend the two principles of justice to us by arguing the value and plausibility of the hypothetical situation that engendered them, Brandt tries to defend Ideal Code Utilitarianism on the basis of its actual consequences. He argues that implementing its principles is the best means to the achievement of our own final ends as well. Whereas for Rawls, the final ends in question were any consistent with the two principles of justice in a well-ordered society, for Brandt, they are the objects of rational desire. As Brandt characterizes them, objects of rational desire are essentially the same as the objects of rational choice described by Rawls's "thick" theory of goodness as rationality (TJ Ch. VII). Like Rawls, Brandt assumes we want the satisfaction of our rational desires, whatever in particular these may be; and he means to argue that the laws and norms of an Ideal Code Utilitarian society are best suited to satisfy them.

But whereas we have just seen that Rawls was, despite his explicit avowals, a Humean disguised as a Kantian, we will now see that Brandt is, despite his explicit avowals, a Kantian disguised as a Humean. On Rawls's view in *A Theory of Justice*, the parties in the original position choose principles of justice on the basis of self-interested desires protected from external influence by certain restrictions on information. On Brandt's view, by contrast, we are to choose principles to govern society on the basis of unrestricted information that causes us to have desires that are by definition rational, whether they are self-interested or not. We have seen in Chapters III and IV that both the restricted- and the full-information models of rational choice yield the trivial result that any such choice qualifies as rational. What most significantly differentiates Brandt's view from Rawls's and calls into question Brandt's self-identification with the Utilitarian tradition is that on Brandt's view, it is what we rationally believe and know, rather than what we simply and self-interestedly want, that is the most important determinant of our choice of social principles:

---

[1] Richard B. Brandt, *A Theory of the Good and the Right* (Oxford: Clarendon Press, 1979). All references to this work are parenthecized in the text.

> [R]ational desire ... can confront, or will even be produced by, awareness of the truth; irrational desire cannot (113).

With this modification, Brandt attempts to escape what we saw to be one of the most central and frequently repeated criticisms of Rawls's metaethics, namely that the conditions defining the original position – self-interest, freedom, equality, the veil of ignorance, the desire to maximize primary goods – were not independently rational but rather already biased towards a certain liberal democratic conception of justice for which no independent justification had been provided. That is, it was objected that Rawls's metaethical methodology was too loaded with value assumptions of its own to provide an objective justification of the well-ordered society it was supposed to support. We saw that this resulted from a more general tension inherent within the Instrumentalist strategy itself, that it can succeed in providing an objective justification only to the extent that it sacrifices the aim of providing a moral justification; and can succeed in providing a moral "justification" only to the extent that it sacrifices the aim of providing an objective justification.

But Brandt's admirable allegiance to the value-neutrality of his conception of rational choice undermines the justificatory force of his Instrumentalist argument for his particular moral theory, just as it did for Rawls. We will see that by stipulating restrictive conditions on motivation and on final ends as part of the rational chooser's psychology at the outset, he robs the derivation of his theory of its rationally persuasive force for us. As did Rawls's derivation, Brandt's also ends up being strictly deductive – but without any pure procedural method comparable to that of wide reflective equilibrium to ensure its soundness.

2. Brandt's Theory of Justification

Brandt means to escape the Instrumentalist dilemma by deploying a weaker, more value-neutral conception of rationality as fully informed choice to provide the metaethical underpinning for his justification of the Ideal Code Utilitarian society (185, 189-193). He means to argue, not that this social arrangement will provide the best resources for satisfying just anyone's desires, but that it will provide the best resources for satisfying anyone's *rational* desires. But he then further qualifies this strategy in a second way, by arguing that this social arrangement will provide the best means for satisfying anyone's *benevolent* rational desires. By restricting the agent's scope of ends and of motivation in these two ways, he makes the agent's resulting choice of something like an Ideal Code-Utilitarian society a foregone conclusion. Of course an agent whose highest priority is the satisfaction of rational desire will choose to live in a society whose moral code is ideally crafted to achieve

this end, other things equal. But this does not answer the question of what society would be best for us, as imperfectly rational agents, to choose. Similarly, a benevolent agent who rationally desires to maximize others' happiness will certainly choose to live in a society whose moral code maximizes everyone's happiness, other things equal. But this does not settle the issue of what society would be best for us as imperfectly moral agents whose motives are not always benevolent.

Given Brandt's commitment to Instrumentalism, there are a couple of strategies open to him: one is to build value-laden normative assumptions into his premises and then "derive" them as conclusions. Another is to build no such assumptions into his criterion of rational desire and so derive none – in which case any desire can be rational depending on the individual and how she responds to facts and logic. We will see that Brandt utilizes both strategies and both fail. But is he doomed to fail? Is it in theory impossible to fashion a procedure or set of formal constraints on rational choice of an Instrumentalist sort, comparable to scientific procedure in science, that is both substantively neutral with respect to content and also yields the right substantive results, at least in easy moral cases? As we have seen in Chapter X.5, hard moral cases are comparable to hard cases in science, in that inferences from a theory for particular cases may be counterintuitive. However, Rawls's conception of wide reflective equilibrium has demonstrated that we can continue to have faith in counterintuitive conclusions if the procedure has proved itself powerful enough to explain and predict commonsense cases (for example, that everyone should contribute to the common good). In order to meet this challenge, Brandt must explicate at a more abstract and formal level what makes the "right substantive results" right in decisions we as readers recognize as such.

For in the end, it is the disinterested and uncommitted reading audience, i.e. we, who need to be convinced. To the extent that Brandt's preferred type of moral "justification" abandons the desideratum of value-neutrality, it thereby seems to abandon the uncommitted audience it was Brandt's original purpose to convince. If we are not already benevolent, i.e. already committed to the project of maximizing welfare, how is Brandt's moral "justification" supposed to justify our choosing a moral system that does? What good does it do us to know that if we were rational and benevolent, we would choose such a system, if in fact we are neither rational nor benevolent?

Brandt does not claim that anyone actually has rational desires. But he does try to convince us that it is in our self-interest to cultivate them. He says,

> I shall ... point to some facts which will recommend rational action and rational desire to everyone or virtually everyone. In other words, I shall cite some *facts, awareness of which will* make the reader more favorably

> disposed towards rational action and desire; *motivate him to some degree to try to do the rational thing;* make him tend to lose interest in goals which are known to be irrational for him, *and to want to conform his desires and aversions to what they would be if he were rational* (151; italics added).

In this passage Brandt describes the justificatory strategy he is going to use with us. In this strategy, awareness of certain facts can motivate us to do the rational thing, and can cause us to want our desires to be rational. We will see that this turns out to be the very same method by which, according to his view, rational desires are fashioned and can be identified. It thereby encapsulates the paradox of Brandt's Kantian type of Humeanism, for it describes a process in which not desire but rather transpersonal rationality, i.e. disinterested, theoretically rational reflection on given facts, is the final determinant of action. So, for example, Brandt's account of rational desires implies that the Harvard professor who merely comes to *recognize* the irrationality of his inclination to decline an offer from UCLA will be sufficiently motivated by this recognition to override his inclination and accept UCLA's offer (125-6).

The strategy Brandt merely outlines in this passage attempts to construct a link between what a fully rational agent would choose and what we ourselves should choose, and to convince us to choose the Ideal Code Utilitarian society because a fully rational chooser would. If Brandt can maintain a genuinely value-neutral conception of rational choice on the one hand and generate from it a specifically moral conception of the just society on the other, he will demonstrate that Instrumentalism as a "pure procedural" methodology can have, after all, the power to generate substantive moral theories, analogously to the way in which untainted scientific methodology claims the power to generate substantive physical theories.

This strategy thus attempts to answer the question of what difference it should make to me to find out that a desire or action is rational, and why I should want to do what is rational. Brandt tries to avoid the time-tested tack on which Gewirth relied, of arguing that to ask why I should do what is rational to do is to ask for reasons; that to ask for reasons is a rational activity; and hence that to ask why I should do what is rational presupposes that I should. Brandt's approach is more ambitious. He means to supply an independent justification of the value of rational desire and action that will demonstrate its intellectual usefulness on the one hand, and move us to be rational on the other. Thus Brandt observes that irrational or uninformed action is not based on knowledge of possible consequences, and we do not like it when our actions have distressing consequences (this is true by definition of "distressing"). It would be nice to be able to foresee and avoid them. Furthermore, he points out, we will be more able to satisfy all our

desires if we act on the basis of full information about them. Moreover, irrational desires interfere with desire-satisfaction generally, as when one has an irrational aversion to going outside but desires a vacation in the Bahamas.

Notice that all three of these reasons appeal to self-interest and the personal comfort of the agent. He says,

> by acting rationally (= avoiding cognitive defects at the moment of decision) we assure that our desires, present and future, are satisfied as fully as possible in the circumstances, irrespective of what they are. By showing that a given action is rational, then, by implication we show that it is chosen by a procedure fitted to maximize satisfaction of desires for the agent. Thus the showing that an action is rational must recommend the act – to any agent, since every agent is an agent with desires. The showing does not accidentally recommend to some people; when we consider that all agents are creatures with desires, it recommends of *necessity* (154).

This is a classic encapsulation of Instrumentalism and of the Humean conception of the self more generally, within which reason is merely a means to the satisfaction of personal desire; and of the Instrumentalist strategy most Humeans adopt. Thus Brandt regards as unproblematic the assumption that self-interested action is always rational. This assumption depends on the prior assumption that self-interested desires are always rational. That is, Brandt assumes that an action's being in our self-interest always provides us with a motivationally effective reason to perform it. We have already found reason to dispute this assumption in Chapter VI, and find more in Volume II, Chapter VIII.

Further considerations that Brandt invokes which recommend rational desire include the fact that we do not like it when our desires and actions are inconsistent with our beliefs. We desire rationally coherent desires:

> The proposal here is that awareness of the fact that one has irrational desires works in a way similar to awareness that one has incoherent beliefs or unjustified fears. One is made uncomfortable by the awareness, and is motivated to remove its source (157).

In line with Brandt's assumption that it is rational to seek personal comfort (159), the desire for rationally coherent desires is itself putatively justified by our desire to avoid the discomforting effect of an awareness that our desires are incoherent. This assumption is in general not sound as a criterion of rationality. But even if it were, it would not furnish a satisfying answer to the question as to the rational status of the desire for coherent desires, and why we should prefer coherent to incoherent desires. That answer, too, will have to await more extensive treatment in Volume II, Chapter VIII.1-2.

Ultimately Brandt's "independent justification" of rationality reduces to a variant on the time-tested tack described earlier. Having adopted an instrumental model of rationality, he then offers an instrumentally rational justification of it. However, this variant is less robust than the traditional one, because one can counter Brandt's defense by simply rejecting Instrumentalism itself as a metaethical strategy. Given the defects already noted, there is independent reason to do this.

3. Desire

Brandt identifies his view as falling within the Humean tradition with respect to the model of motivation he adopts. Revising somewhat his analysis in the seminal article (co-authored with Jaegwon Kim), "Wants as Explanations of Action,"[2] that we scrutinized in Chapter II.1, Brandt defines a desire for something O as follows:

> [A] person 'wants' something O, ... if his central motive state is such that if it were then to occur to him that a certain act of his then would tend to bring O about, his tendency to perform that act would be increased" (26; cf. 30).

Desire, then, is defined by stipulating what an agent would do, were certain conditions to obtain: I want or desire a soufflé if, were it to occur to me that cooking one would get me one, my tendency to cook one would increase.

On Brandt's view, if I desire an end, it is not possible for me think about acting to achieve it without tending more strongly to perform that action. If I can think about acting to achieve it without becoming more strongly inclined to so act, then I cannot be said to desire the object I am envisioning as the outcome of that action. Desire, for Brandt, equals thought plus increased tendency to act.

Thus a desire is a thought-activated physical disposition to act in the service of an end (27, 56). This disposition is activated by the occurrent thought that the act will effect that end. The disposition may be activated without being realized if the thought disposes me more strongly to act without actually causing me to act. In that case, in which the disposition gets stronger but not strong enough to exert sufficient motive influence on my action, the action will not occur (26). The disposition, that is, is a necessary but not sufficient condition, and a contributing cause, of action. Similarly, the action will not occur if the thought does not occur, because the tendency or

[2] Richard Brandt and Jaegwon Kim, "Wants as Explanations of Actions," *The Journal of Philosophy LX* (1963), 425-35; reprinted in N. S. Care and C. Landesman, Eds. *Readings in the Theory of Action* (Bloomington, Ind.: Indiana University Press, 1969), 199-213.

disposition to act has not been activated. So thought is a necessary condition, and the precipitating cause, of action.

Brandt believes that one performs that action which there is the "*strongest net tendency* to perform" (47). So he thinks that no intentional action directed toward a certain outcome will occur without a net tendency to act so as to bring about (or avert) that outcome; i.e. that intentional action directed toward a certain outcome requires a net tendency to act so as to bring about that outcome. This means that in order to perform an act at a certain time, the sum total of the agent's dispositions and traits of character must incline her to perform the act at that time.

On the tautologous interpretation of this claim, Brandt is noting merely that an agent does what she is most strongly inclined to do. This is a variant on the Humean thesis discussed in Chapters II.1 and III, that an agent always does what she most wants to do, and that observation of what she does answers the question of what she most wanted to do. On a nontautologous interpretation, the claim denies that an agent can act against character, or can override her settled habits of mind or behavior in action through the force of some other motive, for example sheer impulse, or reason. On the nontautologous interpretation, observation of what an agent does reveals not only her overriding desire, but thereby her settled traits of character.[3]

Having characterized desire in terms of a tendency (or disposition) to act, and action in terms of the strongest net tendency to act, Brandt then asserts that "no intentional action will occur without desire or aversion directed at it or its outcome..." (66). It may seem that Brandt is here making the same mistake as did Gewirth, by conflating desire and intention. But since Brandt has already defined a desire as a thought-activated tendency to act, this says merely – again – that no action will occur without a prior disposition to perform that action, i.e. that this disposition is a necessary condition – and a contributing cause – of action. On both the tautologous and the nontautologous interpretation of Brandt's thesis, however, the *precipitating* cause of action is not the desire for the end the action effects. It is a component of that desire, namely the occurrent thought that the action will effect this end. Thought, not desire, is what precipitates the performance of action, if anything does.

Brandt's conception of occurrent thought as the precipitating cause of action makes his view a peculiar sort of Humeanism indeed. As we have seen with Nagel, Frankfurt, Williams, Gewirth, and Rawls, Humeans usually stipulate desire as the precipitating conative force in action, and assign

[3] See Brandt's "Traits of Character: A Conceptual Analysis," *American Philosophical Quarterly 7*, 1 (January 1970).

thought and reasoning a subsidiary and instrumental role in achieving its ends. On these views, however, desire itself (not just one of its components) is an occurrent mental event that causes subsequent events, such as thought about how to satisfy it and instrumental action intended to do so. Even on these views, desire is not a sufficient cause of action, since other necessary background conditions also must be satisfied. But once they are, desire precipitates rational calculation and determines whether the act is finally performed or not.

On Brandt's view, by contrast, desire no longer has this centrally determining role. Brandt's modification of the traditional view treats desire as a thought-activated disposition to act rather than as an occurrent event. Of course a disposition can be a contributing cause. But it cannot be a *precipitating* cause because it cannot itself be the thing that activates the disposition. On Brandt's view, thought – i.e. an occurrent belief about the consequence of performing the contemplated action – has this centrally determining role. Without such a belief, on Brandt's view, there would be no increase in an agent's tendency to perform the designated action, and hence no desire to perform it.

When Brandt then maintains that "what an agent does is always a function of his desires at the time; there is no such thing as motivation by beliefs alone" (83), he is maintaining that such a thought is not sufficient in itself, in the absence of a prior disposition, to cause action. He says,

> If some philosophers have thought, as some seem to have done, that a person can do his duty even if so doing is not positively valenced for him [i.e. if he is not disposed to do so], ... perhaps 'out of respect' for duty in some sense, they were wrong; and their psychology of morality needs basic revision (66-67).

On the tautologous reading of Brandt's thesis about desire, there would be no conflict between that thesis and the Kantian claim, which Brandt obviously means to oppose, that one might act out of respect for or knowledge of one's moral duty. That is, it might be true both that one has a net tendency to do what one does, and also that that tendency or disposition itself was caused by respect for or knowledge of the moral law. On the nontautologous reading of Brandt's thesis, acting out of respect for or knowledge of one's moral duty against one's settled character dispositions would be psychologically impossible. Knowledge of one's moral duty could precipitate action only if one had a prior disposition to do it. The extent to which even this nontautologous reading of Brandt's thesis conflicts with the Kantian claim is a matter for debate, since Kant agrees with Aristotle that cultivating the appropriate character dispositions facilitates and is indeed a precondition for

doing one's duty.[4] Nevertheless, this second reading would be the one that seems to underlie Brandt's claim here, as he clearly means to contrast belief and desire as possible causes of action.

But in this case Brandt cannot be supposed to be committed to a concept of desire defined solely as a disposition to action – not even an activated one, because even an activated disposition is merely one of many contributing causes of the actual action. If Brandt means to insist that desire rather than belief or duty is what causes action, then since a disposition to act is merely a contributing rather than a precipitating cause of action, desire itself cannot be merely a tendency or disposition to act. In that case desire itself must be an occurrent event, as the orthodox Humean model of motivation – and the representational analysis of desire I offered in Chapter II.2.1 – assume. Either Brandt must adhere to the orthodox version of the Humean model, which stipulates desire as an occurrent event that precipitates intentional action; or else he is a closet Kantian, for whom occurrent beliefs rather than desire play that role. The following section will adduce further evidence that the latter possibility most accurately describes Brandt's view.

4. Rational Desire

An act, desire, or moral system for Brandt is *rational* if it survives criticism by relevant and available facts and logic (10, 113). It is *objectively rational* if it utilizes all available and relevant information; whereas it is *subjectively rational* if it utilizes all the beliefs rationally supported by evidence the agent has available at the time (72). An *ideally rational agent* is one whose desires and actions are what they would be if the agent had access to and was maximally influenced by all available relevant facts and logic (10, 11). This requires, first, that the agent is vividly and presently aware of every item of relevant information; and second, that the agent's desires have undergone *cognitive psychotherapy*. This Brandt defines as "value-free reflection" on available

[4] Immanuel Kant, *Metaphysik der Sitten, Zweiter Teil: Metaphysische Anfangsgründe der Tugendlehre*, Herausg. Karl Vorländer (Hamburg: Felix Meiner Verlag, 1966), Ak. 399. It is regrettable that Gregor decided to change the translation of *Beschaffenheit* from "disposition" in her original translation of this work to "endowment" in the revised version. Neither is quite accurate, but the original rendering is closer to the literal meaning in this context, which would be something like "habit of character." Kant's subsequent elaboration of this concept as *praedispositio* makes clear that he means to be referring to character dispositions. Compare Immanuel Kant, *The Doctrine of Virtue: Part II of The Metaphysic of Morals*, trans. Mary J. Gregor (Philadelphia: University of Pennsylvania Press, 1971), with Immanuel Kant, *The Metaphysics of Morals*, trans. Mary J. Gregor (New York: Cambridge University Press, 1991).

relevant information. If a desire is not altered by this process, then it is rational for Brandt (113).

Information is *available* if it comprises "the propositions accepted by the science of the agent's day, plus factual propositions justified by publicly accessible evidence (including testimony of others about themselves) and the principles of logic" (13), i.e. if it constitutes the best system of beliefs, justified by publicly accessible evidence, we have at the time of decision (12-13, 112). It is *relevant* if "its presence to awareness would make a difference to the person's tendency to perform a certain act, or to the attractiveness of some prospective outcome to him. Hence it is essentially a causal notion"(12), i.e. if it would influence causally one's tendency to perform the act or one's desire for the state of affairs (12, 112). Relevant information thus includes that which changes one's desires, extinguishes some desires, produces different desires, or simply changes one's behavior. (Brandt further qualifies this definition of relevance so as to exclude the possibility that, for example, performing the multiplication tables before satisfying any desire might discourage all of them, by requiring that the effect of relevant facts or information be local to a particular desire, and a function of its content.)

Brandt's aim is to

> show that some intrinsic desires and aversions would be present in some persons if relevant available information registered fully, that is, if the persons repeatedly represented to themselves, in an ideally vivid way, and at an appropriate time, the available information which is relevant in the sense that it would make a different to desires and aversions (111).

By representing available information in an *ideally vivid* way, Brandt means that the person focuses attention on the information with maximal vividness and detail, and with no hesitation or doubt or stirrings of skepticism about its truth (111-112). Cognitive psychotherapy is thus the process of confronting one's desires with relevant, available and vivid facts and logic.

Brandt believes that desires found to be irrational will extinguish under cognitive psychotherapy, i.e. that once one understands that a desire is based on false or idiosyncratic beliefs and associations, the desire will disappear. If the desire does not extinguish under cognitive psychotherapy, then it is not irrational, regardless of its content (113). The set of desires remaining after cognitive psychotherapy has been undergone will be rational on this account, and so acting to satisfy them will be as well. Thus confronting one's desires with the relevant facts and logic may causally influence motivation, by eliminating as motivational variables those desires that fail to survive the confrontation.

Suppose, for example, that I have a burning desire for a fast-food hamburger. Then I read *Consumer Reports*, from which I learn that fast-food

hamburgers are injected with a poisonous red dye to improve their appearance; infused with hormones to increase their weight, which impede normal functioning of the brain and thereby lower intelligence; that fast-food hamburgers are cooked all at once at 6:00 AM every day, and sprayed with a carcinogenic glaze to retain their freshness; and that while waiting on order, fast-food hamburgers are stored in an insufficiently cooled refrigerator accessible to rodents; and so on. If my desire for a fast-food hamburger wanes the more of this information I receive, then it was not rational to begin with.

But the above passage from page 111 also confirms that the process of cognitive psychotherapy, of fully registering relevant available information in an ideally vivid way, may *produce* rational desires (also see 113). So, for example, one might fully register in an ideally vivid way information about the effect of toxic automobile emissions and develop a desire for better public transportation; or, as actually happened to Attorney General Robert F. Kennedy, fully register in an ideally vivid way facts about the effects of slavery and inner city poverty on the African-American underclass and develop a desire to fight for civil rights. Conversely, one might deliberately avoid developing such desires by choosing to avoid the relevant facts. Cognitive psychotherapy, then, is ultimately a method by which we can cull, shape, and cultivate our settled character dispositions by exposing ourselves and our inclinations to the bracing effects of reality. Relevant and available facts and logic are claimed to have a formative and motivationally influential effect on desires. Brandt's conception of cognitive psychotherapy is thus very close to the concept of a motivated desire which we saw Nagel defend in Chapter VII, and is even closer to the projected effects of the ideal role-taking that we saw in Chapter X.4.3 to be an essential precondition for Habermas' conception of the moral point of view. It reveals Brandt's closet Kantianism most clearly.[5]

Nevertheless, our first question must be whether any of our desires or aversions can qualify as rational on Brandt's intended account; and, if so, how this fact might influence our choice of a moral code. Brandt's conception of cognitive psychotherapy as a corrective for irrational desires also may be understood as a remote heir of Mill's conception of education and experience in cultivating the "higher pleasures" among a society's "informed majority." However, there are two interpretations of Mill's informed majority criterion of

[5] This is not to suggest that Brandt is a *consistent* closet Kantian. He also believes that if the cost of information-gathering can be outweighed by the benefit of acting quickly then it is rational to do so (13, 73). This means that relevant and available facts and logic are dispensable when they conflict with maximizing utility – the standard, neoclassical economic model of rationality that is part of the traditional foundation of the Humean conception of the self.

higher pleasures. On the *democratic* interpretation, any pleasure the informed majority chooses is in virtue of that choice identifiable as a higher pleasure; i.e. the informed majority confers that status on a pleasure in virtue of its choice. On the *elite* interpretation, the informed majority chooses a particular pleasure because their education and experience enable them to recognize it as higher independently of their choice of it. Correspondingly, there are two possible ways of understanding Brandt's cognitive psychotherapy criterion of rational desire. On the *democratic* interpretation, any desire that survives it, for whatever reason, thereby qualifies as rational. On the *elite* interpretation, a particular desire survives cognitive psychotherapy because one recognizes its rationality independently of its psychological survival. That is, confrontation with relevant and available facts and logic enable one to see which desires are in fact rational and which are the result of idiosyncratic associations or warped reasoning. Whereas Mill chooses the elite interpretation of his informed majority criterion of higher pleasures, Brandt chooses the democratic interpretation of his cognitive psychotherapy criterion of rational desires. He thereby avoids linking the rationality of a desire to its particular content, or to the agent's recognition of the rationality of its content. Instead, a desire's rationality is entirely a function of its contingent psychological survival for a particular agent, irrespective of its content. This is a fateful choice that has several unfortunate implications.

First notice that Brandt builds a circularity into his definition of what counts as relevant information in the criticism of intrinsic desire. He suggests that if information causally affects desires and aversions then it is relevant; and that if it is relevant, then it will causally affect desires and aversions. So relevance is a function of causal efficacy exclusively. In an attempt to ensure the value-neutrality of his conception of rational desire, Brandt offers no substantive intellectual criterion by which the content of the information might be assessed for its relevance. But this criterion is then so weak that it cannot generate an identifiably rational choice at all. On this view, it makes no sense to criticize an agent's desires on the grounds that he is not taking seriously information that is relevant to his choices; for if it does not affect those choices it is by definition not relevant, however pertinent it may seem to a third-person observer. And if no information causally affects that desire, then no information is relevant to its rationality. Therefore it is in itself rational for Brandt, regardless of its content or object. So, for example, if my desire for alcoholic oblivion is unaffected by what I have read in *The New England Journal of Medicine* about what alcohol does to the liver or how it shortens one's life, then this information is irrelevant to the rationality of my desire for alcoholic oblivion. If my desire for alcoholic oblivion does not

extinguish when confronted by this information, then it is rational, according to Brandt, no matter how unhealthy or noxious this desire may be.

A second implication of this view would seem to be that either every agent must be supposed to "fully register" available information in the same "ideally vivid" way, which is improbable; or else the rationality of desire is relative to an agent's susceptibility to the effects of cognitive psychotherapy, other things equal. Assuming that agents with full information may nevertheless process that information differently in accordance with their different though fully functioning individual cognitive capacities, it may happen that an arbitrarily selected desire is rational for some of them and not others. The rationality of a desire for a particular agent depends on the contingent and empirical question of how that agent's cognitive capacities happen to respond to cognitive psychotherapy. If an agent is so constituted that the desire extinguishes when confronted with it, then it is irrational for that agent, although it may not be for others. So which desires are rational for which agents will depend on contingencies of their cognitive constitution that are independent of the content of those desires themselves: How strongly facts and logic affect them, and how susceptible to change their desires are in light of them. Brandt accepts this implication (113).

This means that while Brandt's criterion of rational desire may be a universal one, the particular desires that satisfy this criterion cannot be identified as universally rational – nor, therefore, as objectively valid. For any particular desire I may have, whether or not it qualifies as rational for me carries no implication as to whether or not it qualifies as rational *tout court*. To use Nagel's terminology, Brandt's universal criterion of rational desire does not yield an identification of desires that have *objective* value such that anyone would have reason to cultivate them. We saw in Section 2 that it was Brandt's ambition to provide an account of rational desires that recommended them necessarily to every agent (154). But this is precisely what his account makes it impossible to do. Rational desires recommend themselves necessarily only to those particular agents for whom they happen to be rational; and there is no necessity about which ones do. Therefore it is equally impossible for him to demonstrate the objective validity of Ideal Code Utilitarianism as an object of rational desire.

Brandt claims that this account of rational desires is nevertheless value-neutral, in that it does not import into the definition any moral values. Although he cannot demonstrate the objective validity of Ideal Code Utilitarianism as an object of rational desire, according to his account of rational desire, what he can do is try to ground his conception of rational choice in established facts of empirical psychology – about conditioning, information processing, reasoning, and the like – which, he assumes,

themselves contain no hidden, value-laden assumptions. They do, of course. They contain assumptions about the value of the cognitive over the intuitive or emotional; and of self-scrutiny, analysis, and reflection over spontaneous self-expression. Brandt valorizes these skills by assigning them roles as constituents in an ideal of rational deliberation that in turn functions normatively and prescriptively. Still, we have seen in the General Introduction to this project that *this* degree of value-ladenness is a necessary condition of rational dialogue.

Thus Brandt's reductive translation of morally value-laden terms such as "best" and "good" into terms describing the character dispositions of rationality are unobjectionable. He proposes that we understand "the best thing to do" as "the rational thing to do," where "rational" is to be understood in the terms already explained (12, 126-7). Similarly, Brandt maintains, when we say an action is "rational" or "justified," we both describe and recommend it. Brandt's strategy for protecting the value-neutrality of his metaethical program is to defend a type of equivalence relation between normative moral terms and normative psychological terms. This protects the universality of his criterion of rational desires, but it does not increase the objective validity of those desires themselves.

5. Prudence

Should Brandt in fact rule out the possibility that an agent might be motivated to act by beliefs or knowledge alone? Consider, as he does, the case of prudence. Brandt means to locate his view of prudence in direct opposition to that which he takes Thomas Nagel to hold (83-4). On Brandt's official view, I must have a present desire to ensure the satisfaction of what I know my future desires will be in order to ensure it; merely the belief that I will have those future desires is insufficient to motivate me to act.

However, in discussing the rationality of pure time preference, Brandt often slips into the language of belief-motivation. He considers the case in which I now have a desire for a particular satisfaction at a future time *t*, and also the knowledge that at *t* I will also have another, equally strong desire for a second satisfaction at *t*. For example, suppose I now desire to get to bed at 9:00 PM this evening, and also know that at 9:00 PM I will desire equally to watch "Star Trek." About this kind of case Brandt remarks that

> [s]o far my desire now for O [i.e. to get to bed at 9:00 PM], and my future desire for O' [i.e. to watch "Star Trek" at 9:00 PM], appear to come out equally; at least, let us suppose this. *But it is also true that the idea of O now motivates me in a way in which the idea of O' does not. That it does is implied by the fact that I do desire it now.* In view of this fact, does it not seem plausible to say that the total motivation or action-tendency to do what is expected

> to bring about O will be greater than the total motivation or action-tendency to do what is expected to bring about O', at the time? The answer must be affirmative (86; italics added).

Brandt is arguing here, first, that the very fact that I now desire something implies that the idea of that thing motivates me "in a way in which" the idea of something I now know I will desire later does not; and second, that that special motivational oomph I obtain from desiring something now implies that I now desire what I desire *more* than I will desire the thing I know I will desire later.

Both arguments are false. The only way in which my desiring now to go to bed at 9:00 PM need motivate me differently from my desiring at 9:00 PM to watch "Star Trek" is in motivating me now. And the special motivational oomph I receive from desiring now to go to bed at 9:00 PM carries no implication whatsoever that I desire this more than I will desire at 9:00 PM to watch "Star Trek." Brandt's arguments here are based on two unstated assumptions: first, that now desiring O is the same as desiring [that] O [occur] now, an assumption that is violated by any case in which my desire occurs now but the state of affairs I now desire is something I desire to occur later; for example, my present desire to wake up tomorrow morning feeling rested. Thus Brandt presupposes the pure time-preferential assumption he is trying to prove. Second, he assumes that the state of affairs I desire now is somehow sexier or more tempting than the state of affairs I now know I will desire later. But this assumption, too, is false – as shown by the example, in which the more tempting satisfaction is the one I now know I shall desire later. Making plans now to sate or frustrate this future desire, based on and motivated by the knowledge I have now, is what prudence is all about. No present desire is required to motivate me to do so.

But Brandt's analysis of rational desire raises problems even for his official account of prudence as requiring a present desire. The problematic feature of this account is that it is inherently retrospective. That is, it focuses primarily on desires we are already presumed to have, and seeks to modify them in light of facts and reasoning about their origins. As such, it provides a criterion for the evaluation of the desires on which we have acted. In the case in which we subject our known present desires to cognitive psychotherapy, it also provides impetus for reforming those desires in the future. What it does not purport to do is supply an agent *now* with any information that might causally influence *the desires she now first manifests precisely in acting as she does* to ensure the satisfaction of her future desires.

Recall that on the tautologous interpretation of the action-tendency account of desire, we perform that action there is the "strongest net tendency" to perform. The end of the action to which there is the strongest net tendency

is the end we most overridingly desire. Now many alternatives of prudential action are presented to us as possible ends of action only once in our lives. We get to choose only once whether to go right off to graduate school or take a year off after college; only once whether to speak up or remain silent at the moment we first discover a particular injustice; and only once whether to begin saving now at a fixed interest rate of 12% or defer that plan until later when the proffered rates may be lower or higher. In each of these cases, Brandt would say, rationality requires that we examine the cognitive origins of each of our thought-activated action-tendencies in the situation, in order to know which choice we should make. Of course it is particularly important to examine the cognitive origins of that action we have the strongest net tendency to perform, since that is the one that expresses our motivationally overriding desire. However, on Brandt's view, the only way we can know which action we have the strongest net tendency to perform is by performing it; that is how the strongest net tendency to act is identified. This, then, is also the way we identify our motivationally effective desires. The implication of Brandt's action-tendency account of desire – at least on the tautologous interpretation – is that we cannot know that any such newly manifest desire is motivationally overriding in advance of having acted on it. Therefore we cannot correct any such newly manifest desire in light of cognitive psychotherapy in advance of making a decision upon its basis. On the action-tendency account of desire, we may evaluate the rational prudence of many of our decisions only in retrospect.

In Section 2 I argued that a non-tautologous reading of Brandt's act-tendency account of desire commits him to the orthodox Humean account of desire as an occurrent mental event, if he is to retain his allegiance to the Humean conception at all. On this model, he encounters no such problem. If I can know which desire of mine is strongest before I act on it, I can – at least in theory – subject that desire (in addition to all the others) to the scrutiny of cognitive psychotherapy, and make my decision on the basis of its results. In this case the precipitating cause of action will be, not the thought that performing the act will satisfy my desire, but rather the thought that the desire itself is rational and therefore worth satisfying. Reason, to use Kant's terminology, would be an efficient cause of action in such cases. It turns out to be not so easy for Brandt to maintain his allegiance to the Humean conception as it might first seem.

Through exposure to facts and logic I may also produce new desires in the manner described briefly in Section 3. But these newly produced desires will not be thought-activated *dispositions* to act. Such exposure to reality is not the same as the lengthy process of habituation by which character dispositions are gradually fashioned. It is to character habituation as invasive surgery is to

physical therapy. It is literally a moving and often a traumatic experience, an occurrent event that precipitates further events – mental, emotional, conative. It is the kind of experience that does enable one to act in opposition to settled traits of character under the right circumstances – even to act out of respect for the moral law, or to spark a motivationally effective desire to do so. If desires were nothing but thought-activated, settled character dispositions, the experience of exposure to relevant and available facts and logic would be unable to generate them. They would require habituation – in which case they could be activated by thoughts and instrumental reasoning about how a particular action might realize their objects. Exposure to facts and logic would be superfluous.

On the other hand, in those cases in which cognitive psychotherapy *is* sufficient to engender a rational desire, that desire cannot be merely a thought-activated disposition to act, because no process of habituation has formed the relevant disposition (remember that we are considering the non-tautologous interpretation, so it won't do to claim that if we performed the act we must have been so disposed to perform it). Here the desire must be an occurrent event, itself caused by prior deliberation. So Brandt's account of how cognitive psychotherapy produces desires essentially follows Nagel's Kantian model.

So if desires are thought-activated dispositions in the tautologous sense, cognitive psychotherapy can neither evaluate nor improve their rationality. Where cognitive psychotherapy can generate a rational desire, on the other hand, that desire cannot be a thought-activated disposition. If desires are, rather, nontautologous, occurrent mental events of the orthodox Humean sort, then rational actions are precipitated, not by the thought of their instrumentality, but by a Kantian recognition of their rationality.

However, if the orthodox Humean model of desire is, indeed, the one that underlies Brandt's assertion that we must always have a desire to ensure the satisfaction of our future desires in order to do so, it is then unclear from whence the necessity – and hence the universality – of this claim derives. As we have already seen in examining Nagel's view, desires understood in this sense are obviously not the only sorts of occurrent internal events that go on in us. Emotional states such as joy, outrage, or respect, as well as cognitive states such as conviction, doubt, or certainty may be among the others. Nor does Brandt rule out the possibility that these other internal states may have some causal role to play in fueling action as well (89). So it is unclear why desires should be ascribed a necessary part to play in motivating action, and why, other things equal, other internal, occurrent states such as certainty or respect should not at least on some occasions be sufficient.

Indeed, Brandt's account of the motivational efficacy of cognitive psychotherapy would seem to *require* that desire drop out of the motivational chain in some cases. For first, if facts and logic can eliminate some desires from the set of motivationally efficacious variables, then facts and logic can cause me to refrain from performing some actions, namely those caused by the eliminated desires, such that this absence itself may contribute to the satisfaction of my future desires. For example, suppose, at age 20, I know that when I am 60 I will be in great fear of dying from lung cancer, but that 60 is too far away for me to worry too much about that now. If facts and logic can now extinguish my overwhelming present desire to start smoking cigarettes, they thereby can now causally contribute to the satisfaction of my future desire not to die of cancer. In this case the causal variables that explain my refraining from smoking now are facts and logic alone, not desire.

Second, if facts and logic can extinguish all but one present desire, leaving only that one to be motivationally effective, then they play a causal role in generating that one surviving rational desire as motivationally overriding. If facts and logic cause that desire to be motivationally overriding, and that motivationally overriding desire causes me to ensure the satisfaction of my future desires, then again, by transitivity, facts and logic cause me to ensure the satisfaction of my future desires. For example, suppose I know, at age 16, that when I am 45 I will, unless I act now, deeply regret not having finished my high school education now. Also suppose that facts and logic eliminate my desires to do all the things that would interfere with that goal, leaving only a knowledge of my future desire to have finished my high school education now, the future satisfaction of which I therefore now ensure. I then reach the age of 45 and look back with satisfaction and gratitude on my clear thinking and prudential course of action at 16. I thereby satisfy at age 16 my future desire at age 45 to have finished my high school education at age 16. Surely it would be odd to deny that facts and logic had, not only a motivating role, but indeed an overwhelming and necessary motivating role in ensuring the present satisfaction of my future desire to have finished my high school education at age 16, relative to which my present desire to do so played an insignificant role. Indeed, it would not be inaccurate to maintain that facts and logic, rather than my desires, were ultimately causally responsible for ensuring the satisfaction of this future desire. It would seem that Brandt's Kantian conception of cognitive psychotherapy interferes with his allegiance to the Humean model of motivation more often than not.

6. Irrational Desire

Brandt argues that certain actions – hence net action-tendencies, hence occurrent thoughts of the instrumentality of those actions – can be identified

as irrational with reference to the criterion he provides. Which ones, he asks, would not have been performed had the agent had and been maximally influenced by all available and relevant information? Brandt lists several types of such actions; I shall focus on three. First, there are those actions which overlook available options that would have enabled the agent to realize her ends more efficiently, such as walking back and forth three times between the house and, respectively, the supermarket, the hardware store, and the cleaners in order to carry home three separate bundles of goods that could have been transported in one trip by using the car. Second are actions that overlook probable undesirable consequences, such as being too exhausted after the first two hikes between the house and, respectively, the supermarket and the hardware store to then make a third trip to the cleaners.

Some of the problems with Brandt's account of rational desires surface in trying to identify unambiguous examples of either of these first two types. Take the first. If, knowing I had a fully functioning and readily available car, I instead set out to do my errands in three separate walks instead, in what sense can I be said to have *overlooked* the option of driving? Obviously the knowledge that I had a fully functioning and readily available car did not causally affect my net action-tendency. But on Brandt's view, all this shows is that this information was not relevant, and hence that my actions – and so the desires they were intended to satisfy – were rational. In this case taking the car and making one trip instead of three would not have been a more efficient way for me to realize my ends more efficiently because my ends, it seems in retrospect, involved getting some physical exercise rather than discharging my errands as quickly as possible.

Or take the second kind of case, in which the agent overlooks probable undesirable consequences of the action. The same question can be raised here as well. Of course it would have been better if I had managed to complete all three errands rather than becoming too exhausted after the second to embark on the third. But since I chose to walk rather than take the car, we must infer that becoming exhausted after completing my second hike – to the hardware store – is itself preferable to having gotten no exercise at all by taking the car. Once again the inference must be that I did not overlook the undesirable consequences of three hikes over one car ride. Rather, I chose two hikes plus physical exhaustion over one car ride plus three errands done plus no physical exercise. Physical exhaustion was not an undesirable consequence I overlooked but rather a fact that was demonstrably – according to Brandt's criterion – irrelevant to my desires.

The third kind of mistake involves an inability to compute simultaneously all of the rank-ordered subjective probabilities that attach to

each of multiple outcomes of each action alternative. "Evidently," Brandt remarks,

> there is a problem about simultaneous adequate representation of the various features or elements of the alternatives, an inability to get everything adequately, or even equally, before the mind at the same time (76).

Essentially Brandt reasons that we should try to do our best – "try to get the outcomes and their probabilities, as vividly (and equally) before the mind as possible," but that we "necessarily depart from ideal 'rationality' because of this finite capacity of our minds" (78). He quotes R. N. Shepard's observation that "[a]fter a choice of this kind has been made, the decision maker sometimes comes to the realization that this particular choice was not the best even by his own subjective standards."[6] The difficulty is the same as before. Lacking any further criterion of "adequate representation" of the alternatives, we are free to conclude that those elements that failed to inform our decision were therefore demonstrably irrelevant to it. Of course in retrospect, with the benefit of hindsight and the experience of consequences, we may evaluate the alternatives differently. But what was relevant at the time was what causally affected our decision. If information available at that time did not affect it at that time, then it was not relevant at that time.

This general conclusion also applies to the desires and pleasures Brandt claims we would not have if we had all the relevant and available information. Brandt claims that certain beliefs and thoughts that contribute to the formation of desires are incompatible with relevant and available information, and therefore that those desires and pleasures themselves are irrational (89). According to Brandt, we learn to desire things in three ways. First there is *classical conditioning*. Beginning with an innately pleasant (or unpleasant) stimulus to which we respond, that experience is then associated with a neutral experience; we then learn to respond to the originally neutral experience as we did to the innately pleasant stimulus. Then there is *direct conditioning*, in which the thought of an originally pleasant experience is repeatedly paired with an unpleasant stimulus, and so generates withdrawal. Both kinds of conditioning are examples of *contiguity conditioning*, in which we learn to desire something because of its contiguity to something we already desire. Finally, there is what Brandt calls *the principle of stimulus generalization*, i.e. that if we have learned to like a certain thing, we will also like things similar to it, and will like them more the more similar they are.

---

[6] Shepard, R. N., "On Subjectively Optimum Selections Among Multi-Attribute Alternatives," in M. W. Shelley and G. L. Bryan, Eds. *Human Judgments and Optimality* (New York: John Wiley and Sons, 1964), 257-81.

We learn to stop desiring something through deconditioning of three types. First there is *counterconditioning*, a foil to classical conditioning, in which we replace the originally pleasant stimulus with an unpleasant one, and thereby engender the opposite response to the neutral stimulus. Then there is *inhibition*, in which we stop pairing the pleasant stimulus with the neutral experience, until the neutral experience alone eventually stops eliciting the attraction-response. Finally – and most importantly for Brandt's argument, there is *discrimination*, the foil to direct conditioning and stimulus generalization. Here we remind ourselves of the differences between the particular thing we like and other things that are similar but not identical to it (108). So Brandt's basic idea is that a desire is rational if reflecting on its origins, and carefully discriminating it from other similar desires does not destroy it.

There are many particular facts we may reflect on that may change our desires. For one thing, they may be based on false beliefs: for example, suppose I desire to become a stockbroker because I think it will satisfy my parents' wish for my lifelong financial security. I may then discover that they care more about whether I am happy than whether I am financially set for life, and do not think the former requires the latter. Or desires may be the result of misleading or fallacious advertising or cultural bias, i.e. not engendered by actually experiencing the desired or non-desired situation (117). An example would be disliking a low-prestige occupation that may be actually quite satisfying in practice. Or, third, desires may be the result of generalizing from untypical examples, such as assuming, because one was once bitten by a dog, that all dogs are apt to bite. Or they may have been produced by severe early deprivation, as for example the drive to accumulate power may have been caused by childhood experiences of powerlessness; to accumulate wealth by childhood experiences of poverty; or to acquire recognition and visibility by childhood experiences of neglect or marginalization.

Of course it is questionable whether such desires would, in fact, extinguish through reflection on their origins. Does one's desire for a wealthy lifestyle necessarily disappear through reflection on one's early poverty? On the contrary, it may be reinforced by it, as reason for self-congratulation; or be so deeply embedded in one's character that it is for all intents and purposes uneliminable. Brandt's interesting conception of reflection – i.e. of "verbal self-stimulation" – is not sufficiently developed to make this case convincingly. In order to do so, he would need to add some provisos concerning the integration of verbal information with emotion, desire, and memory – perhaps something along the lines of Aristotle's requirement that one have knowledge of both the particular and the universal – in order to ensure the epistemic depth that enables knowledge to actually affect behavior. But even

if he were to do this, Brandt still would seem to be conflating the associative, psychological process of cognitive conditioning with the rational, deliberative process of cognitive intellection. What he offers is not a substantive, philosophical conception that enables us to distinguish rational from irrational desire. Rather, Brandt offers a substantive psychological conception that enables us to distinguish psychologically healthy from unhealthy desire.

Essentially, then, cognitive psychotherapy involves the attempted modification of idiosyncratic childhood associations, learned through contiguity conditioning and stimulus generalization (103), through training in perceptual discrimination and heightening one's awareness of the differences between events formerly associated (107-108). It is the process of searching out and evaluating the empirical accuracy of the associative links between what we instinctively desire, what we are taught to desire, and what we believe. If as infants we associated satisfaction of the instinctive desire for nurture with satisfaction of the learned appetite for condensed milk, we may as adults indulge an overwhelming craving for condensed milk, to the exclusion of friendship and other sources of emotional support. This craving would count as irrational, on Brandt's view, because it mistakenly assumes an exclusive and necessary connection between nurturing and condensed milk, ignoring the role of friendship, affection, and sympathy. Similarly, if we as children associated the desire to play with the thought of parental disapproval or punishment, we may develop into joyless, unhappy and inhibited adults who have an aversion to relaxation and humor. Brandt would regard this aversion to play as similarly irrational, because it falsely presupposes a general rather than a contingent and idiosyncratic connection between play and punishment that leads us to abjure play altogether.

The difficulty this account creates for Brandt is that according to it, just about every conditioned desire may count as irrational.

> [H]ow many responses would turn out to be irrational if one affirmed that any response is irrational if *conceivably*, according to the theory of extinction, it could narrow in its scope or partially extinguish if one repeatedly reminded oneself of the difference between the conditioned and the unconditioned stimulus? A liking for Christmas carols, affection for one's mother; desire for things one likes, desire even for one's own happiness. So, ruling out likes/desires as irrational on so broad a basis would imply the irrationality of virtually all learned likes and desires (144).

Since each particular desire is different in some of its details from others, each can be distinguished from the inductively generalized class to which it belongs. And since each is the consequence of conditioning that is equally

idiosyncratic in some of its details, each is similarly vulnerable to the deconditioning techniques described above. "For this reason," Brandt says,

> I explained 'cognitive psychotherapy' in such a way as to keep some touch with reality; so that a desire or liking ends up as irrational only to the extent that repeated self-stimulations would *actually* diminish it (144-5; my italics).

Brandt's solution to this problem is to distinguish between desires that *would conceivably* extinguish under cognitive psychotherapy, and those that *would actually* extinguish. But this is no distinction at all. By using the subjunctive, Brandt is in the realm of the counterfactually conceivable as opposed to the actual from the outset. His distinction is therefore the distinction between desires D_1 and D_2, such that,

> (1) under certain conditions, I can conceive that {I can conceive D_1 as extinguishing};

and

> (2) under certain conditions, I can conceive D_2 as actually extinguishing.

But if (2) is true, then surely

> (3) under certain conditions, I can conceive that {I can conceive D_2 as actually extinguishing}.

There is then no relevant distinction between D_1 and D_2, since under certain conditions I can conceive both as extinguishing. Brandt's modal operators "conceivably" and "actually" are doing no work here.[7]

By contrast, suppose Brandt's solution had been instead to explain cognitive psychotherapy so that a desire or liking ends up as irrational only to the extent that repeated self-stimulations *will actually* diminish it. A straightforward distinction between the conceivable and the actual would have preserved the desired "touch with reality." But it also would have required him to take a radical empiricist, wait-and-see attitude toward any desire proffered for rational evaluation. This would have made it impossible to speculate in advance of empirical observation on which desires were rational and which were not. We will shortly see that Brandt needs to be able to do this in order to defend his claim that benevolent desires are rational.

[7] In any case, on page 212 he contradicts the distinction he offers here.

Unfortunately, we have already seen in Section 4 that there are independent reasons why no such *a priori* speculation is open to him. According to his criterion, benevolent desires will be rational for some agents but not others. A radically empiricist, wait-and-see attitude toward the rationality of particular desires for particular agents would seem to be required by the Humean program.

7. The Rationality of Benevolence

We have just seen in Section 6 that Brandt's account of inherently irrational desires depended on a behaviorist account of conditioning and deconditioning that itself presupposed the mutability of all desire, both "native" and learned. This distinction is not adequate to the complexity of the relationship between environmental and biological influences on behavior. Nevertheless, Brandt's account works best for learned desire if it works at all. Native desires, on the other hand, would seem automatically to count as rational to the extent that they are sufficiently hard-wired as to be immutable for all practical purposes. Brandt, however, believes it is "logically possible" for these to be changed or extinguished by counterconditioning as well.

We have also already seen that Brandt wants to be able to claim that benevolent desires are rational so that he can then argue the rationality of the Ideal Code Utilitarian society he claims a benevolent chooser would choose. And we have seen in Section 5 that there are obstacles in his account of rational desire that make it impossible for him to do this. Nevertheless, Brandt believes that if he can answer in the affirmative the question as to whether it is rational to be benevolent, he can then reason that since desiring to maximize happiness for everyone is an expression of benevolence, and since Ideal Code Utilitarianism is designed to maximize happiness for everyone, a benevolent chooser would choose the Ideal Code Utilitarian society in which to live. Of course if benevolence cannot be identified as rational, then Brandt's task will be harder. In that case he will have to show that even a self-interested chooser would choose the Ideal Code Utilitarian society. We will see that this is a difficult task indeed.

According to Brandt, a person is *benevolent* if she is (1) intrinsically motivated to produce happiness in others; (2) pleased when others become more happy; and (3) displeased when they become less happy (138). He reasons that if benevolence thus defined does not extinguish under cognitive psychotherapy then it is rational; and that it probably will not extinguish if it is native rather than learned. In defense of treating benevolence as native, Brandt interprets it as a disposition to have empathic and sympathetic

responses to others.[8] He defines an *empathic response* as one in which I respond to another's expression of an emotion by feeling that emotion myself. In a *sympathetic response*, I dislike another's aversive states and like their pleasant states. In a fully rational person, these responses will extend to the anticipated future of the objects of her benevolence, and as much toward future generations as toward the present one. They will also discriminate more finely between rational and irrational distress, and tend to feel empathy and sympathy for rational rather than irrational discomfort (146).

Notice that Brandt's account of empathic and sympathetic responses does not square with his initial definition of benevolence, since one might have empathic or sympathetic responses to others without desiring to maximize their happiness, and might desire to maximize their happiness without having empathic or sympathetic feelings toward them. Empathy and sympathy are emotions, whereas benevolence is a certain kind of desire. There is no necessary connection among them.

Brandt reasons – controversially – that these responses are native because they appear in the second year of life, supposedly before learning can occur. He then argues that benevolence would not extinguish in a fully rational person because the early appearance of sympathy makes it resistant to extinction (143; also 333). This means that whether or not benevolent desires are rational or not is similarly a function of two contingent and empirical conditions: first, whether or not benevolent desires are initially present in an agent – i.e. whether they are instinctive, or, if not, universally learned in the process of upbringing; and second, assuming they exist, whether or not they would extinguish under cognitive psychotherapy or not – i.e. whether they are based on idiosyncratic associations that can be altered through intense and careful exposure to information relevant to them. If there is no exposure to information that would alter them, then they are rational; if there is, then they are not. In the best-case scenario, benevolent desires turn out to have the status of the desire for food or nurturing: they are instinctive and ineliminable by cognitive psychotherapy. In the second-best case scenario, they are at least ineliminable, even if based on learned, idiosyncratic childhood associations. If, on the other hand, cognitive psychotherapy can extinguish benevolent desires in some people, then they cannot be rational for those people. And if cognitive psychotherapy can extinguish benevolent desires in everyone, then they can be rational for no one.

[8] He also says, "or a disposition easily to learn to have empathic and sympathetic responses" (139). But this begs the question, by turning any easily learned desire into a native desire.

Moreover, we have just seen that there is a difference between benevolence and sympathy. So even if it were true that sympathy were resistant to extinction because of its early developmental appearance, this would prove nothing about benevolence. But this is in any case not clearly true. It is a truism that repeated betrayals and exploitative encounters with others may dampen and ultimately extinguish sympathy for others (this is known as "looking out for number one").

Brandt's is an unusual criterion of rationality. Its intuitive appeal is based on the centrality, in more traditional ideal conceptions of theoretical rationality, of reasoning correctly in light of full information. But in Brandt's conception, this traditional ideal conception is subordinated to the empirical question of whether or not such reasoning would have a certain causal outcome or not, such that the desire is rational if it would not and irrational if it would. The oddness of this criterion is brought into relief by the case of benevolent desires, because the question whether or not benevolence is or is not rational is equally traditional and longstanding. Brandt's criterion of rational desire manages to answer the question of whether or not particular benevolent desires are rational without even addressing the question of whether or not benevolence itself is. It seems that the consequence of his attempt to devise a value-neutral criterion of rationality leads him to devise a criterion that is content-neutral as well.

This means that for Brandt there can be no necessary connection between the intuitive rationality of certain final ends as objects of desire and their conformity to or violation of his criterion: both prudence and benevolence might turn out to be irrational in certain cases, if generated by sufficiently idiosyncratic causal chains; whereas howling at the moon might turn out to be rational. A list of content-specific final ends the identification of which as rational would seem a *sine qua non* of any plausible criterion of rationality – prudence, self-interest, friendship, benevolence, justice, etc. – would not serve to test the plausibility of Brandt's. Furthermore, since his criterion of rational desire permits different desires to be rational for different agents, there will be some agents for whom immoral desires are equally rational. This means that there can be no one morally acceptable conception of the good society which all agents would choose as a means to satisfying their rational desires.

So Brandt has not shown that it is rational to be benevolent, his assertions to the contrary. If benevolence is not clearly rational, the desire to maximize others' happiness is not clearly rational. In this case, this desire cannot provide a metaethical justification for choosing a society structured in order to achieve this. As we saw, his strategy was to argue from the rationality of benevolence to the desire to maximize happiness as an expression of benevolence; and

from there to the conclusion that a benevolent rational chooser would choose to live in an Ideal Code Utilitarian society:

> [T]he main inference is quite obvious. For a perfectly benevolent rational person will tend to do whatever will maximize expectable happiness. He will most support that system which as a whole – taking into account probable effects on behavior and the resulting contribution to happiness, and also the costs of the system such as restrictions on individual freedom, the unpleasant pangs of guilt, and the effort of moral education – will maximize the expectable happiness of all sentient creatures. In that sense we can say that he will opt for some kind of 'utilitarian' moral system (217).

Brandt's reasoning here is quintessentially Instrumentalist; and the same problems that beset Rawls's Instrumentalism beset Brandt's. First, the resulting "derivation" is a tautology. Certainly if we assume at the outset that an agent desires to maximize happiness, then, other things equal, she will choose a system that maximizes happiness. In this case Brandt has built into his major premise the conclusion he wants to derive, and the resulting derivation does not, after all, generate a normative moral theory from value-neutral premises. Instead it generates a normative moral theory laden with the same values as were the premises. As we saw in Chapter IX.4.4, Brandt thereby trades an objective justification of Ideal Code Utilitarianism for a moral "justification" that presupposes what it claimed to prove. Brandt must find some other basis on which a fully rational chooser might justifiably choose the Ideal Code Utilitarian society in which to live. Is there one?

We have already seen in Chapter VI that "looking out for number one" is the *un*controversial case of rational action within the Humean conception of the self; and Brandt acknowledges that cognitive psychotherapy may be unable to produce benevolence in such a self-interested[9] person (145). What he presupposes but does not explicitly state is that cognitive psychotherapy may easily produce self-interested desires in a formerly benevolent person: Through maximal and vivid exposure to facts and logic, such a person might, for example, come to revise her sunny conception of human beings as naturally entitled to happiness; and conclude that only those few, herself

[9] Brandt's term is "selfish;" but he must be conflating this with self-interest. A person can be both selfish and also benevolent if she desires both not to share her resources with others and also to produce happiness in them. There is no incompatibility here: if she could satisfy the latter desire without sacrificing any of the former resources she would do it. By contrast, a purely self-interested person would not desire to produce happiness in others unless it promoted her own self-interest in the weak sense defined in Chapter VI.

foremost among them, who have earned happiness deserve it. So whereas benevolence is not clearly rational, self-interest is.

Brandt maintains that even self-interested rational choosers would have reason to chose an Ideal Code Utilitarian society because, first, they would natural prefer some moral code – what he calls a Hobbesian morality – to none at all:

> [T]hey will like the quality of life of a group with a moral system: one with autonomous self-restraint, mutual trust, mutual respect, openness, the absence of need to be on one's guard against malicious or self-serving attacks of any sort. ... Life is more comfortable in such a society, and a completely selfish man will want a society with a moral code on that account (205).

Second, a principle of reciprocity is required in order merely to maximize benefits for oneself:

> If the selfish chooser wants, as he will, protection against crimes against the person, such as assault, negligent injury, and libel, he must choose a moral system which provides the same protection for others, thereby restricting his activities and giving them what they surely want. A selfish person who supports a rule which provides a desired circumstance for all because it, among feasible options, maximizes expectable welfare for him is inadvertently also supporting a rule which will maximize expectable welfare for the group (219).

This reasoning leads Brandt to the following conclusion:

> If we take into account the earlier conclusion that rational selfish persons will support a moral system which provides protections all rational persons want, we have the conclusion that roughly, and in the long run, rational selfish persons will support a happiness-maximizing moral system, not intentionally but inadvertently, since of course each rational selfish person will support his best – his expectable-welfare-maximizing – option among the viable ones open to him (220).

Brandt's conclusion is false. As he acknowledges, a system based on the principle of reciprocity is much less demanding than a happiness-maximizing system (221). A self-interested person who wishes to maximize expectable welfare for himself need not support "a rule which provides a desired circumstance for all." He need support only a much less expensive rule of reciprocity for all that provides the desired circumstance for himself. For example, if he desires to maximize his own wealth, he need not support a rule that maximizes everybody's wealth; a rule that reduces taxes for everyone may well have the effect of maximizing only the wealth of people like him (or, in the best case, only his wealth) because only people like him recoup enough wealth from the tax cut to enable them to pay for the education that trains

them to accumulate more. Where utility-maximization is a zero-sum game, the desirability of a society based on the principle of reciprocity does not imply the desirability of a happiness-maximizing society to a self-interested chooser.

But even if it did, Brandt also acknowledges that the choice would be inadvertent rather than intentional. If it is inadvertent then it has not taken the test of cognitive psychotherapy Brandt proposes as the criterion for the rationality of desire. If the inadvertent choice of an Ideal Code Utilitarian society does not qualify as rational, then it does not matter, for Brandt's metaethical purposes, whether a rational self-interested person would choose it or not. For it does not succeed in justifying that choice to a disinterested and uncommitted reader. We saw in Section 2 that Brandt invoked self-interest to defend rationality to us. We see now, however, that Brandt's appeal to self-interest cannot rescue the rationality of his theory.

So Brandt's metaethical Instrumentalist justification finally fails on four counts. First, we have seen in Section 3 that on Brandt's criterion, rational desires are not objectively justified in virtue of their rationality. Hence even if a benevolent desire for a welfare- or happiness-maximizing Ideal Code Utilitarian society is rational, this does not objectively justify it. But second, we have seen in Section 5 that on his account, benevolence is in any case not rational. Third, we have just seen that Brandt's "derivation" of the Ideal Code Utilitarian society is in any case a tautology that presupposes what it attempts to prove. Finally, it now appears that tinkering with the premises so as to mitigate their value-ladenness effectively subverts the derivation entirely: If the Ideal Code Utilitarian society is not even rational from a self-interested perspective, then it is not a serious candidate for rational justification at all.

8. Cognitive Psychotherapy Reconsidered

We see, then, that the problem lies not with Brandt's conception of the Ideal Code Utilitarian society (at least not in any obvious way), but rather with his conceptions of rationality and justification. We have seen that his conception of rational desire purchases value-neutrality at the expense of viability, for it fails to identify any desire at all as clearly rational. Similarly, Brandt's Instrumentalist strategy of justification encounters the same problems as did Rawls's. The difficulties raised by Brandt's conceptions of rational desire and of justification are, for the most part, mutually independent. But both are supervenient on the Humean conception of the self. It is this, once again, that is the true culprit.

Now I said in Section 4 above that there were two possible ways of understanding Brandt's cognitive psychotherapy criterion of rational desire; and that in choosing the democratic over the elite interpretation, Brandt made

a fateful choice with unfortunate implications. I have just enumerated some of them. What might have been the implications of choosing the elite interpretation instead? Recall that on the democratic interpretation, any desire that survives cognitive therapy, for whatever reason, thereby qualifies as rational; whereas on the elite interpretation, a particular desire survives cognitive psychotherapy only because one recognizes its rationality independently. On this reading, confrontation with relevant and available facts and logic enables one to see which desires are in fact rational and which are warped by idiosyncratic associations or faulty reasoning. On the elite interpretation, it is the rational content of the desire rather than the contingent matter of its psychological embeddedness that ensures its survival.

This interpretation would have required a greater degree of idealization in Brandt's account of cognitive psychotherapy, for it would have required him to ignore the non-ideal case in which the rationality of a desire – for friendly social relations, for example – is obvious, yet we are blinded by other emotions that obstruct our ability to recognize it as such. The elite interpretation also would have required Brandt to say a great deal more about our cognitive capacity to recognize rational content when we are confronted with it, and what it is about us, and about rational content itself, that enables us to do this. In particular, it would have required him to elaborate further the rational capacities we bring to the apprehension of facts and logic, and the rational content of his conception of the Ideal Code Utilitarian society itself. It then would have required him to say what facts about us and it specifically might enable us to recognize its desirability for self-interested as well as benevolent readers. That is, it would have required him to say more about exactly what facts and what logic successful cognitive psychotherapy requires in this case, and how they operate. Just as Brandt's qualified definition of relevant facts and logic was anchored to the content of a particular desire, similarly his elaborated analysis of cognitive psychotherapy as a criterion of rational desire would have been anchored to the content of the particular desire proposed as a candidate for it.

If Brandt had spoken directly to the "fit" between our capacity for independent recognition of the rational and the conception of the good society proposed to be rational, he would have thereby dispensed with the Instrumentalist justification on which he in fact relied; and indeed would have abandoned the Deductivist strategy of which Instrumentalism is an example altogether. Instead he would have lent support to Nagel's Kantian thesis, of attempting to show that our recognition of objectively valid and impersonal reasons can directly motivate action in the absence of desire. And he might have provided indirect support to Kant's own agenda, of arguing that the kingdom of ends is the form of social organization best suited to rational

beings with capacities such as ours. Unless Brandt could have given a competing, equally precise account of our rational capacities such that the Ideal Code Utilitarian society, rather, could have been shown best adapted to them, accepting the elite interpretation of cognitive psychotherapy might have required him to abandon his in any case tenuous commitment to the Humean conception of the self completely. This would not necessarily have been a bad thing.

Chapter XII. Classical Utilitarianism and the Free Rider

In the preceding chapter, I argued that the Instrumentalist reasoning that characterizes the Humean conception of the self could not provide sound justificatory foundations for Richard Brandt's Ideal Code Utilitarianism. In this one, I press that argument to its source in the Instrumentalist, means-end reasoning of Sidgwick's Classical Utilitarianism. His *Methods of Ethics*[1] is perhaps the most rigorous and consistent formulation of that normative moral theory directly and deliberately founded on the Humean conception. Its basic principle is a special case of the utility-maximizing model of rationality, in which the utility to be maximized is social; and its account of agency is founded on the belief-desire model of motivation, in which the desires in question are self-interested, other-directed desires as that concept was explained in Chapter VI.1. Classical Utilitarianism appeals to its stipulated final end – the ideal community in which all members are Act-Utilitarians, to justify not only the basic principle of Utilitarianism, but in addition any and all deviations from it deemed necessary to bring that ideal end-state into existence. I argue here that the Instrumentalist reasoning – and more generally the Humean conception of the self – on which Classical Utilitarianism is based makes it impossible even to articulate coherently a viable alternative conception of the good society, much less to justify it morally. For Classical Utilitarianism, the enterprise of moral justification founders on the theoretical incoherence of the idealized conception of the good society to which the social policies it proposes are supposed to be instrumentally justified as a means.

That we found structurally similarities between the views of Rawls and Brandt should not be surprising, given their shared allegiance to the Humean conception of the self. A similar degree of structural similarity between Social Contract-Theoretic and Utilitarian views can be found at an earlier historical juncture as well. Section 1 grounds Sidgwick's conception of the ideal Act-Utilitarian society in his attempt to solve Hobbes' free rider problem, and contrasts the role in each of the publicity condition, i.e. that the principle on which the agent acts be publicly acknowledged. Section 2 interprets the secrecy stipulation in Sidgwick's proposed solution to the free rider problem as solving a problem left hanging in Mill's; but then draws out the paradoxical implications of this stipulation for Utilitarianism generally. These first two sections call attention to how much Hobbes' Social Contract Theory and Sidgwick's Classical Utilitarianism have in common. The principles that define the Humean conception of the self, together with those that define

[1] Henry Sidgwick, *The Methods of Ethics* (New York: Dover Publications, 1966).

Instrumentalism in moral justification determine a noticeable portion of the content and strategies of reasoning in both theories.

Section 3 shows that in a description of the ideal Utilitarian community that lacks the publicity condition, it is a requirement of social stability that members conceal their adherence to Utilitarianism. Section 4 shows that unfortunately, adding in the publicity condition renders the very conception of such a society incoherent, hence fails to solve the free rider problem even for the ideal case. Section 5 concludes that the failure of the publicity condition for both non-ideal and ideal cases leaves the consistent Utilitarian unable to form psychologically normal and satisfying human relationships, and necessitates a permanent policy of free riding. Hence just as Hobbes' conception of the good society foundered on the problem of the clandestine free rider, Sidgwick's founders on the problem of the public free rider. Both problems are symptoms of the Humean conception of the self they presuppose. A Kantian solution to the free rider problem is offered in Volume II, Chapter IV.8.

1. Hobbes versus Sidgwick on Publicity

All normative moral and political theories in the Anglo-American analytic tradition must confront the challenge of solving the problem of the free rider first introduced by Thomas Hobbes. A free rider is one who enjoys the benefits of others' compliance with a rule but violates it for the sake of personal advantage. If everyone were to behave as the free rider does, the rule would lose its social legitimacy and soon there would be no benefits to enjoy; this form of counterfactual reasoning is the basis of what Onora O'Neill calls Kant's contradiction in conception test.[2] Since anyone can, in fact, reason as does the free rider, the social tolerance of or accommodation to free riding is an incentive to free riders that threatens the continued existence of those benefits for everyone. So the challenge for a normative theory of the good society is to find a solution to the threat of destabilization that the free rider represents.

> Hobbes himself characterizes the free rider as a "fool," who
> hath said in his heart, there is no such things as justice; … seriously alleging, that every man's conservation, and contentment, being committed to his own care, there could be no reason, why every man might not do what he thought conduced thereunto: and therefore also to make, or not make; keep, or not keep covenants, was not against reason,

[2] Onora Nell [née O'Neill], *Acting on Principle: An Essay in Kantian Ethics* (New York: Columbia University Press, 1975), esp. Chapter Five.

> when it conduced to one's benefit. … if it be not against reason, it is not against justice; or else justice is not to be approved for good.[3]

Hobbes' characterization of the free rider's reasoning in *Leviathan* assumes that all individuals are motivated by self-directed self-interest. But the free rider problem does not depend on this assumption. It can be generated as well by agents all of whom are motivated by self-interested but other-directed desires – Rawlsian interests of rather in a self, provided that such other-directed desires have objects with which their agents identify, and that may conflict with one another. So, for example, I may obtain social services for free by secretly cheating on my tax returns, and donate the money thereby saved to anti-war efforts. Or I may fail to pay my annual public library dues, yet make free use of the library as well as the money saved in order to take time off from work to teach an illiterate neighbor how to read. The free rider problem does not depend on egoistic assumptions, as Hobbes assumes. It requires that I take advantage of everyone else's adherence to the rules in order to derive personal benefit by violating them. But it does not require that the personal benefit I derive be a self-directed one.

The basis of the free rider problem is not only the conflict between self-interest and social rules; but in addition the Instrumentalist form of justification that is common to Utilitarianism, Social Contract Theory, and any other normative theory that presupposes the Humean conception of the self (which itself has its origins in Hobbes): If instrumental reasoning justifies the decision to abide by social rules and covenants on the grounds that it maximizes self-interest to do so, then instrumental reasoning equally justifies violating those rules when this maximizes self-interest, whether self- or other-directed.

Knowing this, why does any citizen then ever follow the rules? Why do we not all reason as the free rider does, taking advantage of rules enacted for the social good by breaking them, when we can get away with it, for personal advantage – thereby destroying the purpose and effectiveness of the rules altogether? In a society in which everyone is assumed to be self-interestedly motivated, it is hard to see how social stability can be maintained at all. Hobbes' own solution to this dilemma is to warn that

> he which *declares* he thinks it reason to deceive those that help him, can in reason expect no other means of safety, than what can be had from his own single power. He therefore that breaketh his covenant, *and consequently declareth* that he thinks he may with reason do so, cannot be received into any society, that unites themselves for peace and defense, but by the error of them that receive him; nor when he is received, be

[3] Thomas Hobbes, *Leviathan*, edited by Michael Oakeshott (Macmillan: New York, 1977), Chapter 15, 114. I contrast Hobbes' and Mill's formulations of and solutions to the free rider problem with Edward McClennen's in Volume II, Chapter IV.8.

> retained in it, without seeing the danger of their error; which errors a man cannot reasonably reckon upon as the means of his security: ...[italics added].[4]

In this passage Hobbes argues that by breaking the rules designed to protect all citizens, the free rider thereby announces his unreliability to his fellow citizens, and so absolves them of any obligation to protect him. Since the potential free rider knows in advance that he is thereby undermining his own claim to social protection by breaking the rules, he would be a "fool" to go ahead and do so.

But the force of Hobbes' argument depends on the mistaken assumption that breaking the rules is tantamount to announcing publicly that one is breaking the rules. His argument holds only for the highly unlikely and truly foolish case in which one publicizes one's intent to break the rules for personal advantage, or does so before an audience, or does so but takes no precautions to conceal one's dereliction. To publicize one's intent to free ride in a society in which one's audience does not would, as Hobbes observes, be self-defeating since it would immediately elicit punitive social sanctions. This much seems obvious.

But Hobbes' argument does not address the far more widespread case, in which one takes advantage of others' adherence to the rules by secretly breaking them. Only by being a clandestine free rider can one be a successful free rider. For only in this case can one derive the benefits, gratis, of others' adherence to rules and laws that prescribe, for example, paying one's taxes, voting, financially supporting public radio, etc. Publicizing one's intention to free ride would be to forego these benefits. Thus Hobbes' instrumental reasoning cannot convince a committed, clandestine free rider that her behavior is irrational. If all were to reason as this type of free rider does, there soon would be no stable social rules for anyone to free ride on. On Hobbes' version of Social Contract Theory, clandestine free riders' violations of the law therefore threaten social stability, and are strictly incompatible with the publicity of the principle that justifies them.

All of Hobbes' successors in the Anglo-American analytic tradition attempt to come to grips with Hobbes' free rider dilemma by modifying his assumptions about individual motivation. Sidgwick's solution does so most radically, and therefore fails most dramatically. However, if my conclusions about Instrumentalism in Chapter IX.4.4 are well-taken, a sound solution to the problem requires modification not only in the model of motivation, but also in the corresponding model of rationality; for it is the Instrumentalist form of justification that legitimates free riding as rational and consistent behavior.

---

[4] *Ibid*., 115.

Sidgwick distinguishes between those strategies of action and decision appropriate to Utilitarians living in an ideal social community, and those appropriate in the actual one.[5] In the ideal case, Sidgwick tries to solve the free rider problem by postulating a single, universal other-directed personal motive that motivates and justifies the behavior, including the free rider behavior, of all agents. Instrumental reasoning justifies an individual citizen's decision to break the rules when so doing conduces to social utility, so agents may well break the rules with some frequency. In this case, breaking the rules is quite rightly equated with publicly announcing that one is breaking the rules, for all such violations are publicly accessible. But because all such violations are justified on Utilitarian grounds, social stability remains intact. However, I argue below that this makes not only genuine free riding, but also the consistent and well-ordered social rules on which free riding depends, impossible in theory.

In an ideal community of enlightened Utilitarians, Sidgwick claims, no one would be justified in acting secretly in some way not sanctioned by the accepted moral rules. For even in seeming free-rider cases in which it appeared that one was justified on grounds of utility in exempting oneself from such a rule, this would simply mean that certain qualifications should be added to the rule to cover the exigencies of that type of situation, and thus that these qualifications would apply in all cases relevantly similar to one's own:

> It is evident, that if these reasons are valid for any person, they are valid for all persons; in fact, that they establish the expediency of a new rule...more complicated than the old one; a rule which the Utilitarian, as such, should desire to be universally obeyed.[6] ...If therefore we were all enlightened Utilitarians, it would be impossible for anyone to justify himself in making false statements while admitting it to be inexpedient for persons similarly conditioned to make them; *as he would have no grounds for believing that persons similarly conditioned would act differently from himself*[7] [italics added].

This last clause is ambiguous but significant. On one reading, it would say that a Utilitarian in some situation would expect other Utilitarians to act similarly when "similarly conditioned" because all Utilitarians would react in the same way under some particular set of conditions, that is, that all would reason similarly and thus act similarly. Here "similarly conditioned" would have to mean similar in all respects relevant to the making of one particular decision: similar in personal makeup as well as in circumstances. This is not an unacceptable interpretation of the passage, but it implicitly ascribes to

[5] Sidgwick, *op. cit.* Note 1, Book 4, chap. 5, sec. 3.
[6] *Ibid*., p. 485.
[7] *Ibid*., p. 488.

Sidgwick the view that in a thoroughgoing Utilitarian society everyone is essentially alike, so that a similarity in situation suffices to determine a similarity of response. While this may in fact be a valid implication of the Utilitarian doctrine in its ideal form[8], it is debatable whether Sidgwick would accede to it.

A weaker but more sympathetic reading would construe Sidgwick as meaning that if everyone held to Utilitarian principles, my reasons for acting in a certain way would, in theory, be acknowledged as valid by everyone, even though no one else can, strictly speaking, be conditioned just as I am. Here my supposition that others would behave similarly if similarly conditioned is actually a supposition that, since we all share the same moral principles, others would, if necessary, condone and support my action as being what they would have done if they were, so to speak, in my shoes.

The first reading explains Sidgwick's claim in terms of an assumed uniformity of motives, beliefs, and responses among Utilitarians – not a clearly desirable condition to impose on the ideal society. The second explains it in terms of an implicit acknowledgment of Utilitarian principles as binding on all individuals in the community. The latter would seem more faithful to Sidgwick's intended meaning. He cannot, then, be understood as simply asserting the truism that an exception to a rule that ranges over some class of cases itself ranges over some class of cases. Rather, he is asserting that if everyone justified his actions of grounds of utility, these grounds would be acknowledged under the relevant circumstances as valid and accessible to anyone in any situation – that any action consistently and adequately justified on these grounds could be expected by the agent to receive validation by others in the community. It is in this sense, then, that the rule in question would acquire a qualifying clause, and it is for this reason, seemingly, that Sidgwick sees the principle of Utilitarianism as public in the ideal case. It would seem to be public in the sense that we could not know what someone had done without thereby knowing why; and moreover knowing that they had acted rightly, even in breaking the rule.

This is a consequence of Sidgwick's conception of the ideal community as consisting of what are essentially Act-Utilitarians.[9] Although moral rules are held in common, the decision to follow or not follow them is made on Act-Utilitarian grounds.[10] For even where an apparently Rule-Utilitarian stance is adopted (for example where Sidgwick appraises the utility value of commonsense moral rules), this is done on the grounds that the overall utility of following and promulgating the rule outweighs the personal disutility of

[8] In fact, I suspect that it is, though I will not try to argue this here.

[9] Thus I use the terms "Utilitarianism" and "Act-Utilitarianism" indifferently in discussing Sidgwick's Utilitarianism and its implications.

[10] Sidgwick, *op. cit.*, Note 1, pp. 486-90, *passim*.

doing so. But clearly this does not preclude the case – without begging the question – in which, in the estimation of the Utilitarian, the overall utility of controverting the rule is in fact greater than that of following it.

Here the Instrumentalist calculations are exactly the same for the Act-Utilitarian as for Hobbes' free rider; only the source of utility differs. Hobbes' free rider derives it from the pursuit of self-directed benefit, whereas Sidgwick's Act-Utilitarian derives it from the pursuit of other-directed benefit. In this latter case, it would clearly seem to conflict with Utilitarian first principles to follow the rule.[11] So if everyone follows these rules, this is because all perceive it as instrumentally useful to do so. But because the justification by utility is itself accessible to all members of this community, the very recognition of an agent's situation as being exceptional will determine that an exception should be made, for so would any Utilitarian reason who had access to the facts – including the agent herself.

Thus although the problem of the free rider can arise on the assumption of other-directed as well as self-directed ends so long as the rules are conceived as mere instrumental means to the achievement of those ends, Sidgwick's conception of the ideal Act-Utilitarian society is one in which clandestine free riding is impossible because all agents have the same other-directed end, namely the maximization of social utility. The Utilitarian motive behind any action that violates the rules can be read off from the behavior itself – and thus identify grounds for qualifying the rules in order to include it. In the ideal Utilitarian community, all citizens would reason similarly and conclude, on Act-Utilitarian grounds, that such violations were justified by Utilitarian considerations.

Hence the social instability that widespread clandestine free-riding would threaten would seem not to occur in the ideal Act-Utilitarian community, because rules would be broken only when all citizens could recognize this as maximizing social utility and thus warranting a revision of the rules themselves. Whereas Hobbes' commonwealth is one in which publicity uncovers and ostracizes self-directed deviations from the common good, Sidgwick's ideal Utilitarian community is one in which other-directed violations of the rules are publicly condoned by the universality of the Utilitarian motive itself. Here social instability would seem to be impossible under the two conditions that

(1) all citizens are motivated by the same other-directed desire to maximize social utility, and

[11] For a comprehensive examination of this and related issues, see David Lyons, *Forms and Limits of Utilitarianism* (Oxford: Clarendon Press, 1965), esp. chap. 4C. Also see D.H. Hodgson, *Consequences of Utilitarianism* (Oxford: Clarendon Press, 1967), pp. 3-7; and Peter Singer, "Is Act-Utilitarianism Self-defeating?" *Philosophical Review* 61 (1972): 565.

(2) each publicly breaks social rules when maximizing social utility warrants it.

2. Sidgwick and Mill on Secrecy

Under actual circumstances, however, Sidgwick views the case somewhat differently:

> The Utilitarian may have no doubt that in a community consisting generally of enlightened Utilitarians, these grounds for exceptional ethical treatment would be regarded as valid; still he may ... doubt whether the more refined and complicated rule which recognizes such exceptions is adapted for the community in which he is actually living; and whether the attempt to introduce it is not likely to do more harm by weakening current morality than good by improving its quality.[12]

While the justification for conforming or failing to conform to accepted moral rules is the same for the Utilitarian in the actual as in the ideal community, there is an asymmetry with respect to the accessibility of his principles to others. In the actual, implicitly non-Utilitarian community, the Utilitarian must consider not only the effects of following or not following commonly accepted moral precepts, but also the comparative utility of letting others know the grounds for his decision. For the Utilitarian does not, presumably, do the same things, for the same reasons, as others do in this situation. So whenever his considered actions diverge from those enjoined by the moral rules of the community, the Utilitarian must weigh the utility of this divergence as such, in addition to the utility of the act itself. As Sidgwick argues, the destabilizing effects of this divergence on others may well lead the Utilitarian to conclude that the greatest utility would be served either by performing his action secretly, or by performing it publicly and lying about his reasons for doing so. For in the latter case as well, publicizing the Utilitarian doctrine might undermine general conformity to useful moral precepts even more effectively than his seemingly immoral act, which is at least susceptible to moral or legal sanction. In the non-ideal case, then, Sidgwick's Utilitarian would seem on the face of it to become a Hobbesian clandestine free rider.

So Hobbes' and Sidgwick's views share the following structural similarities. Both rely on the Humean model of rationality. Both rely on the Humean model of motivation. Both deploy the Humean conception of the self in the service of morally justifying comparable conceptions of the social good. And both depend on the same basic Instrumentalist reasoning that justifies clandestine disobedience of social rules: if promoting utility justifies following the rules, then promoting utility equally justifies breaking them.

[12] Sidgwick, *op. cit.*, Note 1, p. 489.

They diverge, however, in the motivational content they ascribe to individual agents' desires and in the particular intentional content of their respective conceptions of utility. For Hobbes the self-interested motivation is self-directed, whereas for Sidgwick it is other-directed. For Hobbes the social good for each agent is the promotion of the satisfaction of that agent's self-directed desires, whereas for Sidgwick the social good for each agent is the promotion of the satisfaction of everyone's self-directed desires (for pleasure, according to Sidgwick).

Hobbes and Sidgwick are also similar, up to a point, in their attitudes toward social instability. In the actual society, social instability for Hobbes would seem to threaten whenever

> (1′) all citizens are motivated to satisfy self-directed desires, and
> (2′) all secretly break social rules.

For in this case, the rules lose their currency and the practices based on them disintegrate. In Sidgwick's actual, non-Utilitarian society, social instability threatens somewhat less whenever

> (1″) all Act-Utilitarian citizens are motivated to satisfy their other-directed desires, and
> (2″) each secretly breaks social rules when maximizing social utility justifies it.

For in this case, the rules lose their currency only for Act-Utilitarians, and the practices based on them are overtly maintained.

Hobbes and Sidgwick thus diverge in their treatment of the publicity condition. Hobbes argues that social instability in the actual society can be minimized if any such free rider is publicly caught in the act, for this tends to call forth a communal, punitive response that discourages free riding. Sidgwick argues that social instability in the actual society is exacerbated if any such Act-Utilitarian is publicly caught in the act, for this undermines not only the Act-Utilitarians' but more generally the community's adherence to commonsense moral rules. So whereas publicity restores a disrupted social order for Hobbes, it undermines social order for Sidgwick.

Each philosopher's argument has equal application to the other's view. It is as true for Sidgwick's non-ideal society as for Hobbes' commonwealth that sanctioning rule-violators may have a deterrent effect on others; and it is as true for Hobbes' commonwealth as for Sidgwick's non-ideal society that publicizing rule-violations recognizes and disseminates an unsavory alternative to adherence that may encourage the very behavior it is intended to deter. One of the problems with Instrumentalist reasoning is that it enables us to advance as a legitimate justification for action any conjectured causal

sequence we like. In the end, both Hobbes' clandestine free rider and Sidgwick's clandestine Act-Utilitarian in the non-ideal, non-Utilitarian society have reason to preserve social stability by concealing their violation of the rules. Only under those conditions can they advance their personal agendas.

Thus on Sidgwick's view, the principle of Utilitarianism should not be propagated at all in its most general form in a non-ideal, non-Utilitarian society, for its effects on the general community may well be subversive of moral conduct if openly acknowledged:

> the opinion that secrecy may render an action right which would not otherwise be so should itself be kept comparatively secret; and similarly it seems expedient that the doctrine that esoteric morality is expedient should itself be kept esoteric .... And thus a Utilitarian may reasonably desire, on Utilitarian principles, that some of his conclusions should be rejected by mankind generally.[13]

Here Sidgwick claims not only the validity of a covert application of Utilitarian principles to support a secret exemption of oneself from some moral precept, but also the validity of secretly adopting these principles themselves. Both are justified on Utilitarian grounds.

So when Mill in *Utilitarianism*[14] dismisses the possibility of such exemption as an objection to Utilitarianism because no doctrine can be formulated which successfully rules it out in all cases, he seems to miss the real point of the objection, which is the unrestricted character of the Utilitarian doctrine: All moral conceptions must admit the possibility of exceptions in practice, but Utilitarianism is unique in rationalizing such exceptions in theory. Mill's own, liberal solution to the free rider problem recommends intensive social conditioning:

> [L]aws and social arrangements should place the happiness or … the interest of every individual as nearly as possible in harmony with the interest of the whole; and … education and opinion, which have so vast a power over human character, should so use that power to establish in the mind of every individual an indissoluble association between his own happiness and the good of the whole … so that *not only he may be unable to conceive the possibility of happiness to himself, consistently with conduct opposed to the general good, but also that a direct impulse to promote the general good may be in every individual one of the habitual motives of action* ... [italics added].[15]

---

[13] *Ibid*., p. 490.

[14] John Stuart Mill, *Utilitarianism*, ed. Samuel Gorovitz (New York: Bobbs-Merrill Co., 1979), chap. 2, par. 25, pp. 24-25. I discuss this passage further in Volume II, Chapter IV.8.

[15] *Ibid*., 17.

Thus Mill's solution attempts to bridge the chasm between Hobbes' actual, imperfect Social Contract-Theoretic society in which free riding is discouraged by the threatened loss of social protection, and Sidgwick's ideal Act-Utilitarian society in which free riding is impossible in principle. Mill's idea is that particular laws, conventions, education, and public opinion are instrumentally justified as a means to instilling motivationally overriding benevolent desires that make self-directed free riding a cognitive and motivational impossibility. However, Mill does not address the question of who is to enact these policies, or how it would be possible to implement them without presupposing the very cognitive and motivational possibilities of free riding it is the goal of these policies to erase. Indeed, it is likely that free riding on the community's adherence to norms of liberty and autonomy would be required in order to instill the relevant benevolent motives.

Sidgwick's formulation can be understood as a response to Mill's silence on this issue. For it spells out and instrumentally justifies the clandestine actions and policies that Mill's solution presupposes. It would seem that these are theoretically implicit in both liberal and classical formulations of Utilitarianism – with all of the paradoxical and alienating implications that follow.

For if the actual community knows that agent *S* is a Utilitarian, they know she justifies all her actions with reference to their utility. And then it is easy for them to infer that *S* will conform to or exempt herself from publicly held moral rules when it maximizes utility to do one or the other. But if the community knows when supporting these rules would not, in *S*'s view, maximize utility, they can infer when she will secretly exempt herself from them. And if they know this, *S* has clearly failed to act secretly, hence she has failed both to maximize utility and to publicly uphold the moral principles of the community. So the Utilitarian must either forego forming the normal human relationships in which her actions and the reasons for them would be known (in degrees varying, say, with the extent of personal involvement in the relationship), which is of questionable utility, or else she must adhere to her Utilitarian convictions covertly on every front. The latter seems to be the more expedient strategy. The Utilitarian cannot, then, make public her convictions without undermining both commonly accepted non-Utilitarian moral precepts and her own attempts to maximize utility in a non-ideal situation.

That the necessity for this thoroughgoing policy of secrecy suggests a difficulty in theory about bridging the gap between the non-ideal and the ideal societies will surely be noted. Unless the Utilitarian is prepared to deny any utility to conforming to non-Utilitarian precepts, it is hard to see what her strategy might be for bringing a community from a non-ideal to an ideal state,

since she cannot, without overall loss of utility, publicize her convictions in the non-ideal one at all.[16]

3. The Pre-Ideal Act-Utilitarian Society

Unfortunately, this problem extends to the so-called ideal society as well. We will now see that even if the Utilitarian could suddenly make everyone else a Utilitarian, without working through the nearly insurmountable obstacles of transition just described, such a community still would not be viable – and for the same kinds of reasons. In Sidgwick's brief adumbration of the ideal community, recall that he says only that everyone is a Utilitarian, hence that everyone justifies both his own and others' actions by the same principle. But we saw that he does not explicitly say that everyone acknowledges this principle publicly. On the sympathetic reading, he seems to assume that this follows from the universal applicability of the principle itself. Now let us look at two variants of the ideal Utilitarian society, one where the publicity condition holds and one where it does not, in order to see whether either alternative will yield the ideal model Sidgwick has in mind.

First let us try to characterize more fully what we might call the "pre-ideal" Utilitarian society, in which the publicity condition does not hold; that is, in which it is not common knowledge that everyone is an Act-Utilitarian. While everyone in fact adopts Act-Utilitarianism as her only rule of conduct (where by "utility" let us understand, roughly, the maximization of happiness, without filling this in any further for the moment), each person does not explicitly recognize others as so doing. Hence although everyone attempts to promote the greatest social utility through her actions, no one views this rule of conduct as the commonly held one. Each person is motivated by benevolence toward the rest of society, but no one is conceived as explicitly sharing these benevolent purposes with anyone else. This is not to say that each conceives the others as selfish and only herself as benevolent. Rather, it is that benevolence is so much an all-pervasive but unarticulated motive of conduct that no one self-consciously conceives of herself or of others in this way. We should try to imagine a situation in which benevolence is so ingrained in behavior that there are no circumstances under which conscious articulation of it is required. It is, let us say, too much of a truism to be worthy of mention. Thus we can think of benevolence in the pre-ideal Act-Utilitarian community as analogous to the motive of self-support in our own. Though we have many reasons and motives for choosing a particular plan of life or vocation, that we should do something with our lives that will insure our own survival in unquestioned – so much so that it rarely figures in an explanation of why we chose as we did.

[16] Sidgwick seems to be sympathetic to this conclusion. See, for example, *op. cit.* Note 1, pp. 474-75, 480-82, 484-86, 489.

The extent of each person's benevolence is, we will suppose, constrained by his own adoption of the Utilitarian doctrine. That is, in estimating the sum of social utility to be achieved by any action, each person counts his own happiness as equal in weight to that of anyone else.[17]

We must also imagine the pre-ideal society, as a variant of the ideal society, to be stable, well ordered, and otherwise successfully operated in the absence of the publicity of its basic social principle of utility. We must, above all, assume it to be reasonably well coordinated: in performing the act with the best consequences, each member takes into account the probable behavior of others and the necessity of insuring against conflicting or self-defeating acts. Thus we can assume, with Sidgwick, that in general the members of this society concur in following certain commonsense moral precepts and rules of thumb on Act-Utilitarian grounds. This is to stipulate that Act-Utilitarian deliberation will, in the absence of the publicity condition, generate these precepts as conventions. Now since everyone is an Act-Utilitarian and hence reasons similarly with regard to the consequences of actions, each person will have no trouble in predicting or assessing the outcome of the behavior of others when deciding what to do. For, although they do not assume that each acts from Utilitarian convictions, they do consider one another's behavior and its overt consequences. To each member of this society, the others behave as if they were Act-Utilitarians in the minimal sense that their actions have, and are recognized by others to have, best consequences under the circumstances.

It is important to emphasize that this state of things does not provide sufficient evidence to any member for thinking that everyone else is an Act-Utilitarian. For to act as if one were is often to adopt *prima facie* non-Utilitarian moral conventions when they have the best consequences – which, in view of the benefits of coordination, will be a good part of the time. This means that one will be unable to distinguish Act-Utilitarians from, for example, highly efficient Intuitionists, on the basis of behavior alone. To identify them as Act-Utilitarians, we must know their intentions and their reasons for acting. But since this pre-ideal society runs smoothly and acceptably in the absence of the publicity condition, members of it will, by hypothesis, rarely be called upon to justify or explain their actions overtly; the practices and conduct of each will mesh harmoniously with those of others. So the opportunity to discover the moral convictions on which they are based will be rare indeed, if not nonexistent. The situation bears comparison with a society in which traditional social roles and practices are, like the benevolent motive, so deeply embedded in the history of the society that talk of reasons and justification for them is otiose. Persons are conceived as inextricably dependent on these roles and practices in a way that practically vitiates the very possibility of calling

[17] I add this proviso in order to avoid the problems inherent in the notion of perfect altruism.

them into question. In this sense, we may say that the pre-ideal Utilitarian society as a whole lacks self-consciousness – not a stringent condition to impose when a society is conceived as functioning smoothly. The roles and practices generated by rational Act-Utilitarian calculation are so embedded in the social structure that their justification is made unnecessary by the smooth and harmonious functioning of the social order itself.

Now if each Utilitarian had no reason to suppose that others shared her convictions, she would obviously have the same good reasons to assume the utility of covert actions in this version of the ideal case as the actual one.[18] Further, she might again correctly assume the greater utility of her esoteric morality than of its public counterpart. This hypothesis is grounded in the supposition that in this society a person's conduct would be fully informed by Instrumentalist Utilitarian reasoning, but would differ from actual behavior only in its degree of efficiency and success in bringing about the best consequences. For Sidgwick, the reasons militating against making public the Utilitarian credo have nothing to do with people's actual relative inefficiency or irrationality, but rather with the damaging consequences of publicly recognizing any rational individual as reasoning in the light of this doctrine; and the further difficulties that would ensue if everyone, or most people, were publicly acknowledged as reasoning similarly. In this respect and under these conditions, the principle of free riding and the principle of Utilitarianism are exactly analogous.

Thus even if our hypothetically placed Utilitarian somehow found out that everyone else were also a Utilitarian, he might well judge even here that it would be better to maintain silence on this point for fear of the destabilizing effects of publicizing it. For note that to say that everyone is an Act-Utilitarian is not obviously to say that they act unanimously, but just to say that each tries[19] to bring about the best overall consequences through his action. And if each has been following commonsense moral precepts in part on the supposition that they reflect the convictions of others and satisfy their valid expectations, we may well expect chaos to result when everyone's assumptions are thus publicly shown to be false. This would seem to hold whether everyone is a Utilitarian or not. So the viability of this variant of the ideal Utilitarian society would, like the non-ideal one, seem to require for its stability the very covert deception of others that Sidgwick wants the ideal society to preclude. An ideal Act-Utilitarian society that excludes the publicity

[18] This seems to be J.J.C. Smart's conclusion as well. See "An Outline of Utilitarian Ethics," in J. J. C. Smart and Bernard Williams, *Utilitarianism: For and Against* (Cambridge: Cambridge University Press, 1973), p. 50.

[19] See Lyon's distinction between accepting and following the dictates of Act-Utilitarianism (*op. cit.* Note 11, pp. 151-52). My reasons for adopting the former, weaker description will shortly become evident.

condition does no better at eradicating the clandestine free rider than a Hobbesian, imperfect Social Contract-Theoretic society that assumes it.

4. Hodgson, Gibbard and Lewis on the Ideal Act-Utilitarian Society

Does an ideal Act-Utilitarian society that includes the publicity condition do any better? Consider the second of the two variants of the ideal Act-Utilitarian society, in which the publicity condition is assumed to hold. I argue here that there is in fact no consistent rendering of such a case. We have mentioned in passing the obstacles that must be overcome in getting from the actual to an ideal Utilitarian society of either kind; and the conclusions of Section 3 suggests that these pains of transition are considerably increased in severity when publicizing the Utilitarian doctrine is made part of the process. But we will now see that even if we suppose this problem solved, the very concept of an ideal Act-Utilitarian society in which the publicity condition holds is incoherent.

This is one way of understanding D. H. Hodgson's argument, which is the basis for his critique of Utilitarianism.[20] Essentially, he argues that truth-telling and promise-keeping would be impossible in a society where all were, and recognized each other as being, Act-Utilitarians. His argument is based on the assumption that one part of the utility of a great many types of social action involves the degree to which they satisfy the justified expectations of another. In an Act-Utilitarian society, no one could have valid expectations about another's actions (to generalize from Hodgson's examples). An agent S would only do an act x if x had the greatest utility; and x would have the greatest utility only if it satisfied the recipient R's expectations. But R would expect x only if R believed that S would do x. Being equally rational, S would know this, hence would do x only if he believed that R expected x. But since S's doing x depends on knowing R's expectations, and R's expectations depend on knowing whether or not S will do x, R has no prior reason to expect S to do x. And since the utility of doing x depends on knowing R's expectations, S cannot determine whether or not doing x has greatest utility. So there is no *prima facie* reason to do x rather than not-x. This dilemma holds for any act x that involves fulfillment or violation of someone else's expectations in its utility index. Notice that this formulation of the problem can be viewed as a consequence of universal free rider behavior that leaves no social convention with any practical normative force.

A sophisticated attempt to refute Hodgson's argument was made by Allan Gibbard in his dissertation.[21] Gibbard interprets Hodgson's argument as claiming that when good consequences depend on the coordination of actions,

[20] Hodgson, *op. cit.* Note 11, Chapter 2.
[21] Allan Gibbard, "Utilitarianisms and Coordinations" (Ph.D. diss., Harvard University, 1971).

rational methods of promoting them are self-defeating unless kept secret.[22] He offers an example of two Act-Utilitarians who, having agreed to play tennis, deliberate about whether to keep their agreement or not, where each will come to the courts if and only if he thinks it sufficiently likely that the other will. Gibbard argues that Hodgson mistakenly infers from this type of instance that what an Act-Utilitarian should do never depends on what he has agreed to do, since he never has sufficient reason for believing that an agreement made with another will be kept. Thus the problem, as Gibbard sees it, is to show that making an agreement in an Act-Utilitarian society under certain circumstances could, after all, alter the expected consequences of the acts open to the two parties – and thus that the agreements that would be kept in such a society are most of those which an Act-Utilitarian would find it desirable to keep.[23]

Gibbard's argument to this effect is strategically similar to David Lewis's in "Utilitarianism and Truthfulness."[24] Both conceive the issue as a limited-alternative coordination problem (Lewis's example is of two rational Act-Utilitarians placed in different rooms who must choose whether to press a red button or a green; only if both push the same button will utility be maximized, and this coordination at best requires their exchanging information about which button each intends to press). Also, both argue that there may be a sufficient condition for solving the problem in an assumption independent of but compatible with the case as stated. For Lewis, it is consistent with the problem to stipulate that the parties will be truthful whenever it is best to instill in the other true beliefs about which one has knowledge.[25] Gibbard's independent premise is the parties' common knowledge of whether their society has kept such agreements in the past, that is, whether they have a history of conventions.[26] Let us look more closely at the latter answer, since Lewis's answer presupposes it as well.

Gibbard's account relies heavily on the analysis of the origin of a convention as the solution to a coordination problem supplied by Lewis in his book, *Convention: A Philosophical Study*.[27] For Lewis, expectations concerning the behavior of other parties in a coordination problem rely largely on precedent, that is, reasoning inductively from relevantly salient solutions to a similar past coordination problem to a most efficacious solution to the present one. The precedent in question may have been established deliberately or by

[22] *Ibid*., p. 156.
[23] *Ibid*., p. 159.
[24] David Lewis, "Utilitarianism and Truthfulness," *Australasian Journal of Philosophy*, vol. 50 (1972).
[25] *Ibid*., P. 18.
[26] Gibbard, *op. cit.* Note 21, p. 164.
[27] David Lewis, *Convention: A Philosophical Study* (Cambridge, Mass: Harvard University Press, 1969), esp. chaps. 1 and 2.

chance, by agreement or tacitly.[28] The more points of similarity between the present coordination problem and its precedent, the more each party is justified in expecting the other parties to concur in solving it in a similar or analogous way. Thus a solution established by precedent gives each party reason to assume a certain pattern of predictable behavior in the others and to calculate his own conduct accordingly.

But I now argue that Hodgson's dilemma implies that no such precedent could be established within the constraints of an ideal Act-Utilitarian society that requires the publicity of its first and only principle of conduct. Successful public adherence to the principle of utility would have to presuppose prior regularities of conduct that could not themselves be justified or determined by that principle. So neither such regularities nor the precedent on which they might be consequent would be forthcoming under conditions that publicly recognized the principle of utility as the only rule of behavior. Thus the choice of action for a consistent Utilitarian would in every case have to be made not between some two alternatives but among a nearly unlimited number of unweighted (or equally weighted) possibilities. I conclude, then, that Gibbard's and Lewis's reliance on the possibility of constructing a weighted, limited-alternative coordination matrix not only fails to address the problem Hodgson raises but indeed begs the question at issue. Absent some better solution to Hodgson's problem, this means that the Humean, utility-maximizing model of rationality is incapable of morally justifying an Act-Utilitarian conception of the good society.

Suppose the alternatives for two rational Act-Utilitarians were in fact, say, between going to the tennis courts and staying at home. This would mean that each party, being fully rational, might simply construct the same coordination matrix, accurately working out the probabilities and the desirable risk for each alternative. Knowing the full Utilitarian rationality of both parties, each would rightly expect this process of reasoning to be fully replicated by the other, so that each would expect the other to arrive at the same solution which she herself did. A condition of this would of course be that each know and assign the same weights and probabilities to each alternative and also have good reason for assuming that the other did so as well. Thus the expectations of each party would, in this case, have to be known to the other: these expectations would derive from their mutually acknowledged rationality and accurate weightings and probability assignments to the positive consequences of each alternative. But because both Gibbard and Lewis give the parties access to the same precedent-setting information, each can justifiably expect the other to assess the alternatives in just this way. The consistent but independent assumptions utilized in Gibbard's and in Lewis's solutions thus amount to *stipulating in advance* the

[28] *Ibid*., pp. 33-36.

weight which each party can be expected to assign, and thereby can expect each other to assign, to the alternative of keeping the agreement to play tennis (or, alternately, of telling the truth about which button one has pressed). But these predetermined expectations themselves presuppose that one can successfully keep agreements or tell the truth in this version of the ideal Utilitarian society. And whether one can or not is, of course, the very question with which Hodgson confronts us.[29]

A refutation of Hodgson's dilemma would, then, have to demonstrate that these practices – and any others that derive part of their utility from satisfying expectations – could be established in an ideal Utilitarian society by Utilitarian reasoning alone, that is, independently of non-Utilitarian expectations. For what Hodgson shows is that if we give no weight to expectations based on the prior existence of truth-telling and promise-keeping, there is no way, consistent with Utilitarian reasoning, of introducing them into deliberation about what to do. Any act that derives part of its utility from the satisfaction of expectations can never be performed, for it will be expected only if it has greatest utility, and it will have greatest utility only if it is expected. But since neither its utility nor whether it is expected can be independently determined, there is no reason for it to be performed at all.

Now one apparent solution to the problem can be found in the plausible assumption that moral precepts or rules of thumb (such as telling the truth and keeping promises) as general guides to conduct have carried over or are remembered from the historically prior non-ideal society. The problem for the members of the ideal society is then simply to publicly incorporate them into the recognized corpus of acceptable Utilitarian behavior. And since, on the Utilitarian account, these rules of thumb have the status they do just because conformity to them usually has best consequences, this should not be too difficult to achieve.[30] Or so it may seem.

However, how are the members of the ideal Act-Utilitarian society to achieve this? To the extent that this program requires them to expressly agree

---

[29] A similar charge can be leveled against the solution proposed by Singer (*op. cit.* Note 11 above), for the basis on which he assumes the positive consequences of truth-telling and promise-keeping to hold in an ideal society seems to me to beg the question in much the same way.

[30] I owe this solution and its elaboration in the second paragraph following to John Rawls, in discussion of the 1975 course paper on which this chapter is based. In his *Political Liberalism*, Rawls states that "any conception of justice that cannot well order a constitutional democracy is inadequate as a democratic conception. This might happen because of the familiar reason that its content renders it self-defeating when it is publicly recognized." Second Edition (New York: Columbia University Press, 1996), 35-36. Conceivably my 1975 paper for his course, and its revised article form ("Utility, Publicity and Manipulation," *Ethics 88*, 3 (April 1978), 189-206) are possible sources of this familiarity.

on the usefulness of these rules of thumb, or publicly stipulate conformity to them under most circumstances, the success of this enterprise once again awaits a solution to the prior question of how they are to agree on anything at all.

But there is an independent problem, even if we suppose that explicit public certification of this kind is unnecessary. It might be, for example, that prior to the full realization of the ideal Utilitarian society, people have been acting on those rules of thumb which generally have best consequences and just naturally continue to act in this manner when all have explicitly become Act-Utilitarians. The difficulty is then generated by the task of combining the notion of rules of thumb for conduct with consistent and public Utilitarian reasoning. For part of what distinguishes the rules of thumb an Act-Utilitarian may consistently adopt (under non-ideal conditions) from those which a putative Rule-Utilitarian adopts[31] is that for an Act-Utilitarian, conformity to the former on any particular occasion requires a decision to do so, based on consideration of whether doing so will have best consequences under those circumstances. But under ideal conditions, the decision to conform to the rule(s) of thumb relevant to the occasion has, and must be publicly recognized to have, the same status as the decision to perform any action, whether it conforms to such a rule or not: an action is to be performed only if it has best consequences. But under conditions in which everyone acknowledges everyone else as adhering to this principle of action, general conformity to rules of thumb can no longer be automatically assumed, for the conditions under which such general conformity had best consequences no longer obtain.

That is, it is no longer true that the non-Utilitarian majority conforms to these rules of thumb, and that violating them would "do more harm by weakening current [non-Utilitarian] morality than good by improving its quality."[32] A community composed only of consistent Act-Utilitarians will conform to such rules only when doing so has best consequences independently of these now irrelevant considerations, and this fact will itself be publicly acknowledged. Consequently, expectations based on previous conformity to rules of thumb under non-ideal conditions must be suspended until it is determined which acts maximize utility under ideal ones. But again, since there is no way of determining the identity of these acts in advance of the expectations aroused by doing them, there is no probability favoring their being performed at all. So it will not do simply to suppose that some act x is expected, then calculate its utility, as Gibbard and Lewis seem to want to do. This is, as Hodgson would say, to engage in mere bootstrap pulling.

[31] "Putative" is used advisedly, in view of Lyon's (*op. cit.* Note 11, above) analysis.
[32] *Op. cit.* Note 1, above.

Thus we can see the significance of Hodgson's (and, for that matter, Kant's) interest in truth-telling and promise-keeping in general as examples for discussion, rather than the instances of these on which Gibbard and Lewis focus. For this reminds us that the question of whether such general conventions of behavior are possible must be settled before the question of whether any particular instances of them are. To the extent that promise-keeping and truth-telling presuppose regularities of speech behavior and shared expectations about how language is used, it is far from clear that communication of any sort would be possible under such "ideal" conditions. To assume in advance that they would begs the question of whether, in an Act-Utilitarian society, such conventions could ever arise, and this is just what Hodgson implicitly denies.

If my account of Hodgson's argument is right, then any attempt to formulate the problem as a choice between two alternatives and to assign weights and probabilities accordingly is bound to fail. For in any confrontation between two Act-Utilitarians, neither can have *any* valid expectations about what the other will do or say, nor any basis for predicting this on probabilistic grounds. This means that the possible alternatives of action open to two Act-Utilitarians in any situation are fairly unlimited. To the extent that doing x depends on its utility, x's utility on whether x is expected, and whether x is expected on x's utility, there can be no sufficient reason for expecting x, for any x, to be done. So the choice is not between, say, going to the tennis courts and staying home, for which a coordination could indeed be established. The choice is rather between going to the courts, staying home, walking the dog, breaking a window, doing a headstand, and the myriad other possibilities that exist between two Act-Utilitarians who have "agreed" to play tennis.

Now sometimes Gibbard seems to talk as though a coordination solution might fortuitously occur and be acknowledged as a solution even if it cannot be expected to occur.[33] Such an occurrence might then provide sufficient reason for expecting it as a solution to future problems. Basically, this is Lewis's argument for the origin of a convention, briefly adumbrated above, and the same objection to it is relevant. If the parties could originally conceive of the issue as a limited-alternative coordination problem, perhaps some such solution to it might be forthcoming. But we have seen that they cannot. Because a basis for expectations about others' behavior is lacking, no act possible under the circumstances has greater initial subjective probability than any other. Hence any act that may be performed cannot be regarded as a solution to the problem of whether to do x or y, since the question of what one should do cannot be made determinate in this way. So even if an act x were the solution to such a problem, the parties would not regard it as such

[33] Gibbard, *op. cit.* Note 21, pp. 167-70.

because they would be unable to conceive the situation in these terms. Because there could be no answer to the question of what act to perform, any act would be equally acceptable.

This is to conclude, finally, *contra* Sidgwick, that the ideal Utilitarian community offers no more of a resolution of the free rider problem than does the actual one. The Utilitarian cannot argue that his esoteric morality is a temporary practical measure, and not intrinsic to the theory in its ideal realization, as Sidgwick wants to do. For in view of the implications of Hodgson's analysis, we seem forced to conclude that secrecy is a necessary ingredient in a viable Act-Utilitarian doctrine. Act-Utilitarian reasoning is a close analogue of clandestine free rider reasoning in all cases, not only in the non-ideal one. And it renders the ideal case impossible in theory. The clandestine Act-Utilitarian free rider is a permanent fixture in the Classical Utilitarian firmament.

5. The Social Utility of Free Riding

Now a Utilitarian may protest that these conclusions are excessively rigoristic. She may claim that the worth of the Utilitarian doctrine is amply demonstrated in circumstances far less radical than those just described. That not everyone could publicly and simultaneously uphold it does not impugn its feasibility under actual circumstances, for its relative success as a consistent and viable moral rule of conduct has served to distinguish it from the spate of seemingly untenable moral views at our disposal. It is, after all, the best we can do. Or so it is claimed.

I now argue that this protest rests on a failure to pursue the psychological implications of consistently adhering to Utilitarianism under actual circumstances in which one is, as in Sidgwick's rendering, a Utilitarian in a non-Utilitarian community. I contend, first, that a sincere commitment to the maximization of utility under these circumstances necessitates a commitment not only to a thoroughgoing policy of secrecy about one's moral views, but thereby to the perpetual covert manipulation of others in the service of one's goals; second, that it is just because ideal conditions can never obtain for the consistent Utilitarian that these morally and psychologically repugnant consequences of his view are in fact unavoidable under actual conditions; and third, that the Utilitarian commitment to maximizing social utility under actual circumstances is therefore just as self-defeating in the non-ideal as in the ideal case.

Assume that I am consistent and fully rational Act-Utilitarian in a non-Utilitarian community, and that I reveal my convictions only to my closest and most trusted friend, who does not share my principles. How can I expect this fact about me to influence our relationship? For one thing, this openness on my part will presumably structure and inform my friend's expectations of me as a Utilitarian, so as to increase the security and affection of our

friendship. But will it? If my friend knows I decide what to do on grounds of utility alone, she will justifiably infer that my openness with her is similarly a matter of policy – that I would not do it if it did not maximize social utility. But my concern with what maximizes social utility clearly transcends the particulars of our relationship; it is this larger goal that I consistently keep in mind and in terms of which the quality of our relationship finds warrant. And if she knows this, she knows my honesty is not merely for the sake of our friendship – not, that is, merely for the sake of my respect and affection for her, but for something in comparison with which the independent value of our friendship pales in significance. Concurring with Sidgwick, I as a Utilitarian "perceive friendship to be an important means to the Utilitarian end."[34] My friend's knowing that I view her in this light hardly seems auspicious for our friendship.

I might, nevertheless, sincerely advise my friend to perform the most socially beneficial, utility-maximizing actions, show her my own consistency in this regard, and demonstrate the good effects that can be brought about. But this will result only if I can somehow convince her that what I tell her to do, and demonstrate by example should be done, is in fact what I want her to do. For while she has no cause to doubt that I see some course of action for her as best, she has no assurance that I think that the best way of getting her to do it is by sincerely advising her to do so. Since she knows that my primary concern is to maximize utility and not to give her my honest opinion, she knows that I might think it best to advise her to do x in order to bring it about that she does y (where y is incompatible with, a side effect of, or part of x). In fact, my friend's suspicion on this point may be justifiably extended to all facets of our interaction: can she ever be sure that my responses to her are not intended to get her to do or think what I think it best for her to do or think, independently of whether she agrees with my judgments? It seems that there is no way of insuring that even the most minimal conditions of moral dialogue are met. As Strawson points out,[35] I may seem to engage her in moral discourse without really doing so. Notice that this is an interpersonal analogue of the instrumentalization dilemma, consequent upon the Humean model of motivation, described in Chapter II.3.2.

In "Freedom, Blame, and Moral Community,"[36] Lawrence Stern argues that Strawson fails to distinguish between calculation and manipulation in his conception of the "objective" attitude. He defines calculation as "subjecting

[34] Sidgwick, *op. cit.* Note 1, p. 437.
[35] P.F. Strawson, "Freedom and Resentment," in *Freedom and Resentment and Other Essays* (London: Methuen & Co., 1974). Strawson's distinction between the objective and the involved attitudes bring out nicely the difference between the perspective a consistent utilitarian must assume and that of most other people.
[36] Lawrence Stern, "Freedom, Blame, and Moral Community," *Journal of Philosophy* 71 (1974): 72-84.

whatever feelings one has to the constraints of policy, to getting the result one is aiming at," while manipulation is "subverting or bypassing another person's rational or moral capacities for the sake of some result."[37] In therapy, for example, he points out that although calculation must enter into the attitude of the therapist toward the patient insofar as the patient's well-being is the result being aimed at, manipulation need not, since the therapist can make full use of the patient's rational or moral capacities in furthering this goal.

Although the distinction is well taken, it is important to see how closely intertwined these two must be in the attitude of a consistent Utilitarian toward anyone else. Here, calculation implies manipulation. For in order to promote the result the Utilitarian is aiming at, namely, maximizing social utility, it will be necessary to bypass the other person's rational and moral capacities just in case publicly acknowledged agreement on the goal to be achieved is lacking – which, as we have seen, must be true for the consistent Utilitarian in all cases. For example, the Utilitarian may enter into a friendship for reasons of utility, as Sidgewick suggests; but if the other person enters into it solely because he likes and respects the Utilitarian personally, and the Utilitarian knows this, it is unlikely that the latter will succeed in bringing about a commitment to the relationship from the former except by manipulation, by getting the other person to commit himself, without openly presenting her Utilitarian calculations of how to best maximize utility as a reason for doing so. On the other hand, cases in which calculation would not necessarily imply manipulation are just those, for example, business relationships, in which people are consciously committed to cooperation in some enterprise the goals of which are mutually acknowledged. But since mutual acknowledgment and cooperation in the goals of Utilitarianism have been shown to lead to insuperable difficulties, the implication holds in this case.

The possibility – indeed, the necessity – of a consistent policy of manipulating others and calculating their responses as variables in the service of a larger goal reveals a serious problem with the very concept of a consistent Utilitarian doctrine, quite aside from the difficulties discussed so far. As we saw at the outset, the first principle of Utilitarianism is a special case of the nonmoral utility-maximizing model of rationality, in which the particular utility to be maximized is general social utility. Now normally, the utility-maximizing principle is called into use under circumstances that themselves determine whether or not the question, "Does this act conduce to *G*?" is relevant; for most goals are such that not all actions, and not all circumstances, will obviously bear on their realization. For instance, if I wish to learn horseback riding, my taste for foreign films will not be a relevant

[37] *Ibid*., p. 74.

consideration in any obvious way. The restricted nature of the goal *G* itself places certain practical, probabilistic constraints on the class of actions that are to be assessed for their expediency in bringing it about. Hence *G* will not form some part of the purpose of every action an agent considers.

Compare the principle of utility. Where *G* is "general social utility," this can be further fleshed out in any number of ways. What is important to note is that any more specific formula substituted for it (e.g., "everyone's well-being," "the general level of happiness," "the satisfaction of everyone's wants," etc.) must be sufficiently general so as not to rule out the happiness, pleasure, or satisfaction of wants (however these terms are then suitably defined) in advance for some particular person, for in Sidgwick's Classical Utilitarianism, this would be to decrease the total sum of utility. But since there is no prior way of determining what makes every person happy or satisfied, or what constitutes a person's pleasure, there is no prior way of eliminating from consideration any purposeful action whatever as a possible means to this goal. The goal of maximizing social utility is so encompassing that any act performed in an interpersonal context must be evaluated for whether its consequences are relevant to, or constitutive of, its realization.

This means that a concern with social utility must form some part of the motivation of a consistent Utilitarian in any interaction he engages in, indeed in any plan of action he undertakes: this is the full sense in which Utilitarianism provides the only rules of conduct for one committed to this doctrine. It may be that some such activities are then found or judged to be irrelevant to furthering social utility. But this can only be a consequence, and not a presupposition, of an evaluation to which every action is initially susceptible. This reveals the extent to which calculation – hence manipulation – must inform the Utilitarian's every decision, action, and deliberate response.

So if people know that someone is a committed Utilitarian, they are bound to feel somewhat used or manipulated, somewhat suspicious of her manifestations of feeling, involvement with, or professed regard for them, and somewhat resentful of her attitude. Clearly, it is more expedient for the Act-Utilitarian to keep her moral convictions to herself, both in this world and in the ideal one. But because all such convictions must be clandestine, the consistent Utilitarian must, in all of his interactions, free ride on the moral transparency and candor practiced by those around him.

If a Utilitarian adopts this policy of thoroughgoing secrecy about his principles, there is nothing to prevent his doing a great deal of good. In fact, it is entirely possible that he may do better than most of us, for his actions will be more fully informed by rational Utilitarian deliberation. But he will stand in a unique and not wholly desirable relationship to everyone else in the world, whether or not they share his convictions. He will, as it were, have to keep his own counsel on every occasion. He will be unable fully to confide any of his plans, hopes, or intentions to others, or to reciprocate in

attachments and dependencies on them, insofar as these involve shared commitment and trust; and he will be unable to seek or find confirmation of criticism of them in the convictions of others.

It is questionable how worthwhile a Utilitarian might then find his doctrine. For not only would it seem to necessitate a degree of alienation from others, the psychological cost of which cannot be repaid. It also requires a rather strong, and probably incorrect, assumption about human psychology if the agent's hierarchy of values is to be stable. It needs, that is, to assume that a person's convictions can thrive on purely internal support, that a lack of confirmation and esteem of these beliefs by others will not erode or weaken their importance and value in the Utilitarian's own mind. This is not to claim that our deepest convictions require public consensus in order to reassure us of their validity; it is just to question the sense in which moral principles can be believed to be correct if they are in principle acknowledgeable only to oneself.

This has certain consequences, implicit in the discussion above, for how the consistent Utilitarian must regard other people. She must, without confiding in them, both do what she sees as best promoting general utility, and also do what is necessary to get others to do the same. The telling asymmetry of justification we mentioned near the beginning of this discussion thus reappears in a stronger form: the Utilitarian acts from well-reasoned motives that accord with her deepest convictions, while she requires and expects no such deliberation on the part of others. It is sufficient for her purposes that they perform the (from her perspective) requisite actions and have the requisite thoughts and responses. But however complex or reflective these may be, they will have no independent validity for the Utilitarian. She accords them importance only insofar as they coincide with her plan. That is, she views the opinions, feelings, and deliberations of other people – indeed, other people themselves – as instrumental to her goals.

Now in our dealings with young children, we often get them to do or think things that are instrumental to worthwhile goals we have for them, by arranging their environments in certain ways; by dissembling, simplifying, or ambiguating the facts in answer to their queries; by carefully selecting the states of affairs, behavior, and utterances to which they shall be privy. We rightly justify these practices by pointing out a child's malleability and the necessity of paying close attention to formative influences during the years of growth. This filtering of influences is necessary, we point out, if children are ever to reach a sufficient degree of maturity and inner stability to understand and cope with the complexities and perils of the world from which we now seek to shield them. Thus a child's eventual competence, maturity, and autonomy adequately justify our covert practice of manipulating his environment. Such a practice is rightly held to be ultimately in the child's best interests.

Like the Utilitarian, parents have reasons of utility for not disclosing some of their intentions and beliefs to their children: they would be disruptive, misunderstood, have an untoward effect on psychological development, and so on. Like the Utilitarian, parents cannot require their children to make a considered judgment or mature confirmation of the validity of these beliefs. For this reason, parents, like the Utilitarian, can have a satisfying and affectionate relationship with their children, but do not expect to form the same complex relationship of mutual affect, trust, dependence, and respect that is possible with a friend or equal. Like the Utilitarian, the morally best act for a parent is often the one with the most favored consequences for others, that is, the children: parents often feel that their beliefs and efforts will be sufficiently vindicated if only their children grow up to be happy, mature, and productive adults who have recognized the value of their parents' strategies and have developed a minimal gratitude for their efforts. These are the worthwhile goals in the service of which they manipulate their children.

But at this point the analogy with the Utilitarian importantly fails. For we have seen that there is, after all, *no* future state of things with reference to which the Utilitarian free rider may justify his policy of secrecy and manipulation and in the light of which this policy might eventually be dispensed with and commonly validated, in retrospect, as a means to the worthwhile goal of moral maturity. That is, there is no point at which the attitude of the Utilitarian toward the rest of the community can develop past the paternal attitude of a parent toward his child and no point at which the Utilitarian can eventually share with others a relationship of mutually acknowledged respect as mature, autonomous, moral adults. The consistent Utilitarian, then, must permanently regard himself as though he were the only adult in a community of children.

Now multiply this pervasively unhappy attitude by the number of consistent Utilitarians, and calculate the social utility thus maximized accordingly. The more Utilitarians there are, the less social utility is maximized and the more widespread the unhappiness and self-censoring alienation inherent in simply being a consistent Utilitarian becomes. Conversely, the fewer consistent Utilitarians there are, the healthier one's social relationships, the happier one consequently feels, and the more efficiently social utility can be maximized. Social utility can be most efficiently maximized, it would seem, by abandoning the commitment to Utilitarianism entirely.

Chapter XIII. Baier's Hume

In this volume so far, I have examined the metaethical views of a wide range of twentieth century Humeans, to whom Hume's own views are of varying degrees of interest; as well as the Humean conception of the self these views all presuppose. Neither these views nor their presuppositions have purported to represent Hume's actual philosophical theses accurately, either in part or in whole. Rather, Hume's views have been positioned as the inspiration, the historical authority, and hence the legitimating imprimatur for the more contemporary philosophical views that purport to reconstruct them.

Yet Hume's own views are often cited as refutation of some of the arguments I have so far advanced. It is often protested that these criticisms do not apply to Hume himself, whose philosophy is much more nuanced and complex; and hence that they merely target a straw man – or, at the very least, wrongly incriminate Hume through guilt by association. So it is now time to take the measure of these protests, by edging toward a direct, exegetical confrontation with Hume's philosophical position itself. In this chapter I target one of Hume's most well-known and highly regarded advocates; one who purports explicitly to model her own nuanced and complex philosophical views on Hume's – and who thus offers exegetical insight into Hume's views in the very act of formulating her own. In the next chapter I examine Hume's own arguments independently, in order to ascertain whether they provide the warrant that Annette Baier, as well as other, less historically-minded contemporary Humeans claim they do.

In *Moral Prejudices,*[1] Baier argues that Hume's view constitutes a radical alternative to what she views as the predominant Kantian, Social Contract-theoretic paradigm in normative moral philosophy. On Hume's view, she claims, we must assign highest priority to such "thick" moral concepts as caring, trust, and familial and gender relations. Correspondingly, we must ignore or reject the traditionally more abstract concerns of moral philosophy, such as justice, obligation, and freedom. Baier's account of Hume's own conception of the self is notable for its complexity and refusal of reductionism as a value in theory-construction. Hence it reaches well beyond the bare bones utility-maximizing model of rationality and the belief-desire model of motivation appropriated into contemporary metaethics from it. Nevertheless, I argue here that, just as Kant incorporated Hume's insights into a yet broader and more subtle conception of the self, Baier's own defense of Hume similarly presupposes the very Kantian conception of the self she purports to reject.

---

[1] *Moral Prejudices*, (Cambridge, Mass.: Harvard University Press, 1994). Page references to this work are parenthecized in the text preceded by "MP."

Section 1 describes and evaluates Baier's distinctive Anti-Rationalist methodology in philosophical argument, what I call the indexical approach; and compares it with that of Hume himself. Section 2 examines Baier's attempt to replace Social Contract Theory with an analysis of familial and power relations among mutually dependent and morally unequal and imperfect human agents. Section 3 traces Baier's Humean critique of Kant's model of Social Contract Theory and attempt to install Hume as a better role model for contemporary moral philosophy. Section 4 evaluates Baier's arguments, and concludes that, just as Hume's own theses often presupposed the foundational assumptions that Kant made explicit, Baier's case for Humean moral philosophy presupposes the Kantian assumptions it is supposed to replace. Section 5 reconstructs Baier's Humean analysis of trust as the foundation of her normative moral theory. Section 6 evaluates it with respect to several criteria a *bona fide* moral theory must satisfy, and concludes that Baier's moral theory satisfies them – without, however, dispensing with the background, Social Contract-Theoretic moral assumptions she attacks. Section 7 evaluates Baier's indexical approach to philosophical exposition, and suggests that the radical and unwarranted extent to which she deploys it marks the point of departure between her own philosophical priorities on the one side, and those of Hume, Kant, and Socrates on the other.

1. Baier's Humeanism

Baier states her allegiance to Hume's own ethics and epistemology up front. The care and sensitivity with which she treats Hume's texts confirm this. Baier calls our attention to very many of Hume's little known or previously disregarded arguments and pronouncements – for example, his account of familial relationships (MP 69), of single motherhood (MP 73), and his thesis that knowledge acquisition depends on the structure of governmental authority (MP 90). She situates these accounts in the contexts in which their significance becomes clear; and she draws forth their implication and applications to contemporary issues of concern. Moreover, Baier's insistence on situating Hume's moral philosophy in the context of his epistemology and psychology,[2] even though it would ease her own task to treat it in isolation, is exemplary and unusual. She tolerates undogmatically Hume's philosophical flaws, and is willing not only to defend his insights and contributions, but to take him to task, in print, however reluctantly, for inconsistencies, howlers, or morally, politically or intellectually indefensible claims – his condescension toward women, for example (MP 52); or his racism (MP 291). That is, Baier treats Hume not as an authority to be propped up at

[2] Baier develops this argument at length in her *Progress of Sentiments: Reflections on Hume's Treatise* (Cambridge, Mass.: Harvard University Press, 1991). Page references to this work are parenthecized in the text preceded by "PS."

all costs, by dogmatically suppressing or trying to rationalize even his most absurd or idiosyncratic pronouncements; but rather as a valued friend, with whom she is in deep and continual dialogue of the sort that yields revelations of the self, the other, and the world, and with whom careful scrutiny and criticism of particular claims, arguments, or choices of words is an expression of respect. Ironically, given her Anti-Rationalist views, her approach to Hume seems at first glance a paradigm example of Socratic metaethics, practiced across time and cultures between two refined philosophical intellects. Part of the pleasure of following Baier's engagement with Hume's ideas is in her ability to ignite in us the same abiding interest in what Hume actually has to say, whether it serves her own philosophical agenda or not, that fuels her own enduring friendship with him. In the end, I argue, Baier badly betrays Hume's own commitment to Socratic metaethics, and so the foundations of her attachment to his thought. But even this betrayal is enacted with the intention of extending Hume's approach into a radically new style of philosophical dialogue.

The approach to topics of interest to her that Baier shares with Hume – and fellow travelers Michael Slote and Bernard Williams – consists in forswearing system and abstraction for an associative and concrete approach to her topic. I shall call this the *indexical approach*. The indexical approach is itself a practical application of Baier's more generalized, philosophical distrust of theory (MP 194). We have already seen in examining Williams' views about ground projects and Slote's views about pure time preference that there are inherent paradoxes in adopting the indexical approach as a matter of principle. But never mind. By the *indexical approach*, I mean a way of writing and arguing philosophy that (1) equates the order of exposition with the order of thought; and (2) equates the object of exposition with the objects of experience. I now say something about each of these in turn.

1.1. The Order of Exposition and the Order of Thought

Some philosophers try to organize their thoughts for their readership by linking them in ways we learn as part of our philosophical training: conceptually, logically, or dialectically. This is *the order of exposition*. Thoughts are linked *conceptually* if the exposition proceeds primarily through the conceptual analysis of terms, ideas, and theses – which analyses contain further key terms, ideas and theses that are in turn subjected to analysis, and so on until the topic has been fully mapped. Much of Harry Frankfurt's work can be described in this way. Thoughts are linked *logically* if the exposition proceeds largely through the exploration of the logical relation – entailment, disjunction, analogy, contradiction – between premises. Judith Jarvis Thomson's work often lends itself to such a characterization. Thoughts are linked *dialectically* if the exposition proceeds largely through the statement, criticism and its rebuttal, and further refinement of a thesis. Alan Gewirth's

work often satisfies this description. Of course most philosophers, including the ones I have named, make use of all of these expositional devices, and more, to some extent, in different proportions.

All of these devices have in common the imposition of system upon a set of thoughts that may not, and often do not, naturally occur in that order. The natural *order of thought* for many philosophers is associative, nonlinear, and dependent on factors – environmental, perhaps, or temperamental, or chemical – that we do not fully understand and cannot fully control; and the task of whipping them into expository shape is more or less onerous to that extent. One of the reasons why Hume's *Treatise* is such a difficult read for some of us (and why Baier's mastery of it is therefore so impressive) is that, aside from a few unenthusiastic gesturings in the direction of system in his exposition – for example, his basic division of the text into sections on epistemology, psychology, and ethics; or his adoption of the basic empiricist taxonomy of mental contents (ideas, impressions, relations of ideas, etc.), Hume is primarily an associative and naturally linear thinker who treats topics as they occur to him in connection with those he has already treated. Sometimes this leads him to saw off the branch on which he is sitting, as, for example, when, after having demolished the concepts of causality, substance, space, time, and the self, it occurs to him that he then has to explain how it is that objects seem to persist even when our perception of them is discontinuous.[3] His exposition here resembles a particularly savage granny's knot, and is just about as much fun to untangle. What saves him for the reader partial to system is that he is a seductively good writer.

Baier's writing has many of these features, too. As was true for Slote, Baier's approach to the issues of interest to her lends itself well to the essay form, because she simply addresses a very loosely connected succession of topics as they come to interest her for different reasons – invitations to write on particular topics for collections or conferences, personal experience, her interest in Hume, her rejection of what others have said, etc., and trusts (rightly, for the most part) that they will add up to a coherent whole. "[T]opics [do] somehow seem to present themselves," she says,

> and to have some sort of link with topics I [have] addressed at some earlier point. ... Many of [my] essays were originally prepared as lectures for particular audiences, often at conferences with themes chosen by the organizers or as essays for anthologies on a particular theme, so whatever unity they have certainly does not come from ... logical step-by-step progression ..." (MP xi-xii)

[3] "Of scepticism with regard to the senses," Book I, Part IV, Section II of *A Treatise of Human Nature*, Ed. L. A. Selby-Bigge (Oxford: Clarendon Press, 1968), 187-218. Page references to this work are parenthecized in the text preceded by "T."

Nevertheless the unified whole Baier's essays add up to is delineated by the distinctive character of her philosophical sensibility; and partly definitive of that sensibility is her complete disinclination to impose "logical step-by-step progression," or even to seek it. By contrast, some philosophers (me among them) choose what to write about, when, and in what form, on the basis of prior, systematic constraints on what their theory requires.

1.2. The Object of Exposition and the Object of Experience

Like Hume, Baier also tends to equate the object of exposition with the objects of experience. For she shares with Hume certain philosophical interests of an experientially concrete and pragmatic nature. Hume's skepticism resulted from a refusal to credit any source, whether epistemic or ethical, as authoritative, beyond the evidence of his own experience, and this resulted in an attack on almost everything. Like Hume, Baier is an iconoclast at heart. And as is true for Hume, her targeted icons include – but are not exhausted by – the abstract edifice of rigorous metaethical speculation that, in her view, obscures the centrality to human morality of such "thick" ethical concepts as love, trust, vulnerability, contempt, responsibility, violence, relationship, and family. Like Hume, Baier means to remind us of our fleshy embodiedness, our biological and social interdependence, and our fallibility. Like Hume, she means to resituate our moral dialogues and deliberations in the context of the pervasive and incorrigible uncertainties – about ourselves, our relationships to others, and what will happen next – that define our actual existence.

Baier also targets the impersonal style of abstract philosophical analysis that fundamentally characterizes the discipline of philosophy. For it she often substitutes anecdotes, narratives, and personal remarks, with less benign results. In this enterprise she is faithful to the equation of the object of exposition with the object of experience. The experiences are often her own – of reactions she has had to others, stories she has heard, events with which she has first-hand familiarity. But in this regard she decisively parts company with Hume himself, whose style, though associative in just the manner the substantive content of his interests would seem to require, is also relentlessly impersonal, analytical, and abstract. I argue in Section 7 below that this divergence has fatal consequences for the plausibility of Baier's analysis.

1.3. Baier's Project

Despite – or perhaps because of – Baier's similarities with Hume, her Humeanism is original. For her familiarity and comfort with detailed scrutiny of the historical texts enables her to avoid the embarrassing and all-too-pervasive spectacle in which ignorance of the history of philosophy leads to repeated reinventions of the wheel. To her very great credit, Baier renders unto Hume what is Hume's; and this frees her to make her own bid, not

merely for greater appreciation for a Humean perspective, but for a shift in our conception of the concerns that centrally define the enterprise of normative ethics. So Baier's Humeanism is unlike those of the philosophers already considered, in that it embraces rather than avoids the challenge and messiness of exegetical confrontation with Hume himself.

Essentially, Baier's bid has two parts, a substantive one and a critical one. The substantive part proposes to replace the analysis of moral obligation as the centerpiece of normative ethics with an analysis of power relations. The subtlety and detail of this substantive part of her project make her difficult to typecast as a consequentialist, deontologist, or virtue theorist (those taxonomic refuges of metaethicists who have nothing normative to say[4]). The critical part of the project proposes to replace what Baier views as the dominant authority of a rigidly masculine and sexist Kantian approach to normative ethics with what she views as a more sensitive, pragmatic and feminist Humean approach. In some places, Baier's criticisms are compelling, and require us to rethink the scope and enterprise of normative ethics. But in others, her treatment of her adversaries is so careless and disrespectful both of her chosen targets and of the enterprise of – yes, abstract – philosophical analysis that it is almost impossible to read without losing one's temper, particularly if one is a Kantian. I start by examining the substance of her objections to the prevailing tradition, and proceed from there to discussion of the alternative model she advocates, leaving my very strong reservations about some of the methods she deploys to the last section of this chapter.

2. Baier's Critique of Social Contract Theory

Shifting our focus from obligation to power relations means replacing the analysis and derivation of prescription with description, strategizing and problem-solving as our highest normative priorities. There is, in fact, a great deal more description than prescription in Baier's *Moral Prejudices* – more so than readers of moral philosophy books may be used to. She is a keen observer of what we actually do, and how we negotiate moral relationships both personally and institutionally; and her explicit attachment to personal anecdote and autobiography (MP 194) (for which she unnecessarily asks our forbearance (MP xii)) should be seen in this context. Her sometimes gory accounts of what actually happens – both to her and to others, often in the academic context – under the rubric of moral transaction lends first-personal authority to her generalizations; and is a necessary antidote to the naive and self-aggrandizing assumptions about human nature that ground so much contemporary moral philosophy and thus restrict it either to the realm of theory or to the genre of motivational, you-can-do-it-if-you-try literature. Some respond to this lack of realism with respect to our actual limitations and

[4] This remark is not merely a dig but also a summary of my arguments in Chapter V.

flaws by arguing that we must revise downward the values expressed in normative moral theory so as to accommodate them.[5] Baier's consistently indexical approach to the problem leads her simply to present us with the facts.

This should not be taken to imply that no moral ideals or convictions stand behind her approach; quite the contrary. But these ideals and convictions function more as pragmatic and pluralistic goals to be achieved through action of which we are realistically capable, than as impossibly abstract and removed standards for the dispensation of praise and blame. When Baier makes recommendations – she almost never prescribes – about how we should address a particular moral dereliction or issue, one senses that she tests their viability against the capabilities and limits of her own moral agency. The self-knowledge, humility, and moral engagement these recommendations display have far greater authority than the commands, coercion, and threats she repeatedly deplores. But we shall see that the question of who actually issues such commands, coercion or threats surfaces repeatedly.

Baier's most far-reaching recommendation is that we need to replace the prevailing Social Contract-Theoretic model of moral interaction between agents who are conceived as free, equal, and autonomous with a familial model of moral interaction between agents who are mutually dependent, unequal in power, and connected by material, social, and biological necessity (MP 120). Her critique of the Contract-Theoretic model dismantles its most basic constitutive concepts, namely those of obligation, the nature of individuality, and the presuppositions of equality, freedom of choice, rationality, and rights.

She begins by attacking the notion of *obligation* as the central concept of ethics. Her objections to this notion are threefold. First, take the foundational obligation of *promise-keeping*. This presupposes that one has been raised to take promises seriously; and this, in turn, that we have the obligation to raise our children to be morally competent promisors. But this requires that we raise our children lovingly, and "an obligation to love, in the strong sense needed, would be an embarrassment to the theorist, given most accepted versions of 'ought implies can'" (MP 5). For the most part, traditional moral theories of the modern era have nothing to say about proper child-raising,

[5] See, for example, Bernard Williams, *Ethics and the Limits of Philosophy* (Cambridge, Mass.: Harvard University Press, 1985); Samuel Scheffler, *The Rejection of Consequentialism* (Oxford: Clarendon Press, 1982); and Susan Wolf, "Moral Saints," *The Journal of Philosophy LXXIX*, 8 (August 1982), 419-438. I do not think it plausible to interpret these and other similarly inclined authors at face value, as arguing against moral theory altogether, for reasons I have touched in Chapter I and develop at length in Volume II.

and so nothing to say about how to insure the stability of a community's morality over time. On Baier's view, the observation of moral obligations and natural duties that characterize a morally competent agent presuppose the primary virtue of being a loving parent (MP 6). The importance of child-rearing, a responsibility traditionally allocated to women, has been overlooked, presupposed and exploited by traditional moral theory (MP 9).

Second, when we compare the contractual obligation to keep one's promise with the range of actual situations we may encounter, about which we can make no such contractual provisions – the terrorist plumber who fixes my drains but blows up my house, the subversive teacher who imparts the history of ethics but might turn a son against his parents – it becomes clear that this obligation is a special case not suited to be the foundational or definitive one (MP 116-7).

Third, to be under an obligation in general implies that one is unfree, so if obligation is the central concept of ethics, then the justification of coercion is the central problem (MP 4, 12). The Social Contract-Theoretic tradition has concentrated on cases in which punitive enforcers of obligation are entrusted with more power than the obligated agent, but this neglects that large variety of obligations and expectations of others that fail to conform to what Baier calls the "coercive model" (MP 13). Moreover, the main support for the stability of a moral code cannot be threats and coercion, on pain of infinite regress. Rather, it must be – again – the loving parents we trust to raise trustworthy moral agents (MP 14).

From obligation Baier moves on to the traditional contract-theoretic conception of *individuality*. Following Carol Gilligan, Baier rejects Rawls's conception of individuality as defined by one's rational life-plan on three grounds. First, it presupposes an obligation of noninterference by others that devalues the importance of close association with others (MP 24-5, 119). Second, it ignores the necessity of care – and therefore interference – by others in the lives of the relatively powerless, in order to avoid their neglect, isolation and alienation (MP 24-5, 29). And third, it permits the exploitation of those – traditionally women – who voluntarily choose to follow a more communal and care-based conception of individual growth. As Baier observes, "[a]s long as women could be got to assume responsibility for the care of home and children and to train their children to continue the sexist system, the liberal morality could continue to be the official morality, by turning its eyes away from the contribution made by those it excluded. The long unnoticed moral proletariat were the domestic workers, mostly female" (MP 25-6).

Baier's attack on the Contract-Theoretic conception of *equality* is essentially that it levels and disregards the actual power imbalances that characterize our relationships – between parents and children, states and citizens, doctors and patients, large states and small among them. By

pretending that these imbalances do not exist in order to bring about their eradication, she argues, we obscure from ourselves the true nature of our moral relationships to others who are inferior or superior to us in power; and so we often do violence to ourselves as well as to them (MP 28).

About the Social Contract-Theoretic conception of *freedom of choice* Baier is equally pessimistic. She points out that moral relationships to others – for example, of child to parent, or of later to earlier generation, are not always freely chosen. Shifting the moral emphasis from contractual justice to care, she argues, "goes with a recognition of the often unchosen nature of responsibilities of those who give care, both of children who care for their aged or infirm parents and of parents who care for the children they in fact have" (MP 30). I think what she means here is that elevating care and concern for others to a central moral role, rather than simultaneously presupposing and devaluing it, as she claims the contract-theoretic tradition does, assigns priority to the moral quality of the social relationships we actually have, rather than to a conception of ourselves as moral agents who can create the ones we want *ab nuovo*.

Baier also levels a frontal attack on the *rationalism* she finds inherent in the Contract-Theoretic moral tradition, i.e. on the "assumption that we need not worry what passions persons have as long as their rational wills can control them" (MP 30). If unequal power relations – in particular those between parent and child – are the moral norm, rather than contractual relations among equals, then the primacy of rationality must be re-evaluated accordingly (in this conviction I am, obviously, squarely in her camp). Rational control may be a necessary condition for adequate parenting, and maybe even sufficient in a distant father-figure who has no substantial involvement in day-to-day child-rearing. But "primary parents [usually mothers] need to love their children, not just to control their irritation" (MP 31).

Finally, Baier's critique of the Contract-Theoretic notion of *rights* begins with her acknowledgement of the fundamental connection between rights and speech. To speak at all, she observes, to assert or claim anything, is implicitly to assert or claim the universal right to speak and the right to be heard. "We are a right-claiming and right-recognizing species," she says, "and these claims have a built-in potential for contested universalization" (MP 224-5). If Baier is correct in maintaining that language expresses respect for persons (MP 232, fn. 5), and that rights language expresses our emerging self-consciousness as individuals and the externalization of our individual powers (MP 232-3; also cf. 236-8, 240-1), then rights would seem to have a foundational role that conflicts with Baier's agenda to displace the Contract-Theoretic model from ethical primacy.

But rights, on Baier's view, are not only individual but also individualistic. She finds an inverse correlation between our increasing

tendency to claim more universal rights, and our decreasing tendency both to beg or feel grateful for the granting of those benefits to which we believe ourselves entitled, and to respond generously to those who do (MP 226). Historically, she reminds us, such claims by some group of the oppressed against some class of oppressors were exclusive in nature: of commoners by the aristocracy against the king in the Magna Carta; of women and slaves by white men against the king in the U.S. Bill of Rights. "Slaves, nonlandowners, the poor, women, criminals, children, nonhuman animals have all been out-groups in relation to those in-groups who have claimed their rights in famous proclamations and manifestos" (MP 226), and rights can be manipulated to shore up traditional privileges as well as to extend them to the disadvantaged (MP 228-9, 231). Because rights are inherently capable of conflict, different rights-claiming groups de-emphasize or circumscribe different rights in order to claim others as universal (MP 228, 230); and coherence within a list of universal rights is purchased at the cost of precision and sometimes even presence of specifiable content (MP 228).

So on the one hand, rights are often seen as more inherently individual and basic than obligations, duties, or responsibilities (MP 237-8). But on the other, Baier argues, the preservation of particular individual rights under particular circumstances are merely the tip of the iceberg of morality that are "supported by the submerged floating mass of cooperatively discharged responsibilities and socially divided labor" (MP 241); and the rules of discussion that must be observed in order for a speaker to be heard demonstrate this (MP 241-2). The individuality of rights is a foundational social fact about language-producing creatures. But the individualism of rights reveals it to be a cooperative social enterprise and so not as foundational as first appears.

3. Baier's Case for Hume over Kant

Baier observes that since Hume preceded Kant, we know what Kant thought about Hume, but not what Hume might have thought about Kant (MP 268). She proposes to remedy that asymmetry, by mounting a Humean critique of these defects of the Kantian Social Contract-Theoretic model. She thinks these defects, and many others, can be remedied by turning from Kant to Hume as a source of guidance for fashioning a contemporary, realistic, and fully responsive moral philosophy. As a beleaguered Kantian, I find odd her assumption that Kant provides the dominant model for contemporary moral philosophy, and have at least suggested the pervasiveness of the Humean model in preceding chapters. But of course an aggrieved sense of philosophical deprivation is not the exclusive preserve of either camp.

Begin with Kantian Rationalism. Here Baier suggests we substitute Hume's concept of reflexion. She characterizes reflexion as a "response to a response ... a sentiment directed on sentiments" (MP 72, 81-3); and

impressions of reflexion as "feeling responses to how we take our situation to be" (MP 131). Baier thinks Hume's sentiment-based concept of reflection performs the same function as Kant's concept of reason (MP 72), but without its coercive and individualistic connotations (MP 62). Reflection and premeditation will "make a difference to the operation of natural motives and passions" (MP 66), just as Kantian reason is purported to do. Like Kantian reason, it separates mature and self-critical moral agents from mere conformers (MP 72). And Humean reflection carries with it the same potential for self-deception, i.e. the tendency to overlook or dramatize our moral derelictions in ways that obscure our effective moral beliefs (MP 67). Moreover, there is a developmental progression in levels of reflection, from a child's instinctive feeling of sympathy for another, to our more considered sympathy for another's resentment at insult or injury, to the "reflexive turning of these capacities for sympathy, for self-definition, and for conflict-recognition onto themselves, to see if they can 'bear their own survey'" (MP 72). Humean reflection, Baier concludes, thus can provide an account of moral development analogous and in some ways superior to the prevailing Kant-Piaget-Kohlberg model.

Finally, Humean reflection may play the same authoritative role in deliberation as reason: Hume's reduction of rational inference as traditionally understood to custom and habit is coupled with the stipulation that those customs and habits of thought are authoritative that survive the test of reflection (MP 81). This Baier interprets as meaning, first, that we must be able to continue the custom or habit in question even after we have "thought long and hard about its nature, its sources, its costs and consequences" (MP 81); and second, "we must be able to turn the habit in question on itself and find that it can 'bear its own survey'" (MP 81-2).

The most authoritative survey is that which is administered by the passions, including socially dependent ones, on all the operations of the mind traditionally identified with reason – understanding, memory, demonstration, causal inference, and the assumptions of physical and mental continuity. Of these the passions ask, "'Would we perish and go to ruin if we broke this habit? Do we prefer people to have this habit of mind, and how important do we on reflection judge it that they have it?'" (MP 82). Baier defends this method of reflective survey as our only resource for evaluating our values. "We ... can do no more," she argues, "to revise lower-level evaluations, than to repeat our evaluative operations at ever higher, more informed, and more reflective levels" (MP 84). Humean reflection, according to Baier, thus trumps Kantian reason on two counts: first, Kantian reason itself is nothing more than custom and habit; and second, it, like all such habit must be subjected to higher-order authoritative evaluation by directing the passions upon it.

Interestingly, Baier's advocacy of Humean reflection as a better substitute for reason represents a revision from her earlier identification of reason with

Humean belief in *A Progress of Sentiments.* There she drew heavily on Hume's remarks at T 118-120 about the causal influence of belief on the will, passions, and action, in order to resist the implications of Hume's infamous claim at T 415 that "[r]eason is, and out only to be the slave of the passions, and can never pretend to any other office than to serve and obey them." Baier argued as follows:

> Belief and obsessive or even passing thoughts influence our passions and motives, and so influence our will. Hume takes this point to be already established long before he gets to his most famous and infamous Book Two claims about the impotence of "reason alone" to produce or prevent any action.
>
> He begins the resumed discussion by claiming that popular and traditional philosophical talk of 'the combat of passion and reason' is strictly nonsense (T 413-415). Since passions incorporate the influence of reason, since they presuppose beliefs, they would be in combat with themselves if they resisted the influence of belief (159).

The final sentence of the above citation is not to be found in Hume's *Treatise*. Rather, it is Baier's own reasoning in defense of Hume's claim in the preceding sentence. In the following chapter I suggest how very much at odds it is with the other, surrounding claims Hume makes in that section. But for present purposes it is sufficient to note Baier's (1) implicit recognition of the distinction between the passions on the one hand and reason/belief on the other; and (2) seeming equation of belief "and obsessive or even passing thoughts" with that reason which Hume argues repeatedly is reducible to custom and habit. In Baier's *Progress of Sentiments*, then, the fundamental distinction is between the passions, and the habits of thought we identify as reason and/or belief.

However, Hume's characterization of belief as itself "*an act of the sensitive, [rather] than of the cogitative part of our natures*" (T 183; italics in text) may have led Baier to rethink this seeming equation. If that which "influence[s] our passions and motives, and so influence our will" (PS 159) is "an act of the sensitive [rather] than of the cogitative part of our natures" (T 183), whereas "reason … exerts itself without producing any sensible emotion" (T 417), then clearly it is the sensitive part of our natures rather than reason that influences our passions, motives and will for Hume; and "belief and obsessive or even passing thoughts" may, as sensitive, enter into the "reflective survey" (MP 81-2) conducted by the passions upon all of our habits and customs. Thus in *Moral Prejudices*, belief and the passions are on the same, "sensitive" and motivationally effective side of our natures, whereas reason is on the "cogitative" and motivationally impotent side. The passions administer the most authoritative survey of our customs and habits because only the passions, guided by their sensitive beliefs, can change them.

Baier thinks a Humean ethics has, in addition, several further advantages over a Kantian one. First,

> Hume's ethics, unlike Kant's, make morality a matter not of obedience to universal law but of cultivating the character traits which give a person 'inward peace of mind, consciousness of integrity,' and at the same time make that person good company to other persons. ... To become a good fellow-person one doesn't consult some book of rules; one cultivates one's capacity for sympathy .... Hume's ethics ... does not reduce morality to rule following (MP 54-55).

Second, Hume is more of a cultural and historical relativist than Kant. Whereas the latter regards reason as the source of moral rules and views them as universal, Hume regards them as generated by self-interest, instrumental reason, custom, tradition, chance and human preference, and as varying among different communities (MP 55). Baier regards this as self-evidently in Hume's favor, since it conforms more closely to the facts of human social life.

Third, Humean ethics does not ascribe priority to relationships among autonomous equals, as the Kantian Contract-Theoretic model does (MP 60). Rather, Hume assigns a central role to the family, by beginning his analysis of cooperation with that within the family between parents and children, who are unequal and dependent; and with what he describes as "the strongest tie the mind is capable of" (T 362), i.e. that between parents and children. In being intimate, unchosen, and between unequals, this relationship stands in sharp contrast to the Kantian Contract-Theoretic model (MP 60-61). Baier faults the latter on this score, for its emphasis on self-chosen commitments and consequent inability to account for the "duties both of young children to their unchosen parents, to whom no binding commitments have been made, and of initially involuntary parents to their children" (MP 61).

Fourth, whereas the aim of a Kantian Contract-Theoretic morality is freedom, and the problem how to achieve it given conflict with others who want it equally, the aim of Hume's morality is to solve the "deeper ... problem of contradiction, conflict, and instability in any one person's desires, over time, as well as conflict among persons" (MP 61). Conflict internal to the agent and conflict among agents are mutually interactive; and this, Baier thinks, further highlights the contrast with a Kantian ethics that conceives moral agents as independent, autonomous, and distanced.

Baier devotes a great deal of attention to the relative cruelty of Kant's and Hume's ethics, though she declines to say what she means by the term "cruelty" (MP 269). Her target is the comparative pressures that are exerted on moral agents to get them to conform to the moral norms in question, and the punitive measures that are taken when they fail to do so (MP 268). Whereas

Kant's is what she and Allan Gibbard[6] would call a guilt morality that redirects others' anger at our faults through the attribution, feeling, and punishment of guilt, Hume's is what Baier describes as a shame morality (MP 270), in which moral motivation is inspired through approbation of virtue, the sense of virtue, and "also the principles, from when it is deriv'd. So that nothing is presented on any side, but what is laudable and good" (T 619). So instead of feeling guilty for doing what is wrong, one feels shame for not measuring up to "what is laudable and good." Here anger at injury, hatred, and the need for revenge is, as for Kant, expressed through law enforcement; but we attempt to prevent such feelings from "rising up to cruelty," "the most detested of all vices" for Hume, by expressing our approbation of those in whom these passions are reflectively controlled, i.e. by hatred for cruelty, for excessive hatred, for revenge that is bloodthirsty and out of control (MP 271).

Baier points out that Hume frequently appeals "to what we would be gratified or 'mortified' to have others discern in us, to what 'renders a man either an object of esteem and affection or of hatred and contempt" (E 138/174).[7] Other terms of evaluation Hume frequently uses are "'honorable and shameful, lovely and odious, noble and despicable " (E 173/214), as well as aversion and disgust (MP 281). And Hume seems to rely on the shaming practices of public scorn and social ostracism to instill the relevant virtues (MP 282). But, Baier argues, this may sound worse in theory than it could possibly be in practice, since "[w]e cannot *all* be shunned, and even if we can all be occasionally derided, this must be merely occasional. Unless just a few of us display any Humean vices in a high degree, then we simply cannot make avoidance our normal response to perceived vice in others" (MP 282). Thus unlike Kantian morality, which is expressed in "stern imperatives, demanding direct obedience," Hume's is "morality with a light touch," expressed in the optative mood, as "a list of welcome characteristics, with enforced commands carefully limited to where they are needed and known to be effective in controlling 'incommodious' passions or in protecting others from their fallout" (MP 289).

Kant's almost unconditional approval of *Lex Talionis* as a principle of just punishment – public humiliation for the public humiliator, death for convicted murderers, castration of rapists and pederasts, and total ostracism

[6] Allan Gibbard, *Wise Choices, Apt Feelings* (Cambridge, Mass.: Harvard University Press, 1990), pp. 297-298.

[7] David Hume, *Enquiry Concerning the Human Understanding and Concerning the Principles of Morals*, Ed. L. A. Selby-Bigge, Second Edition (Oxford: Clarendon Press, 1966). Paragraph/page references to this work are parenthecized in the text preceded by "E."

from human society for those guilty of bestiality,[8] grates on the sensibility that employs "laughter more than hectoring commands" (MP 289), and for this Baier takes him to task. "One wonders," she comments, "about the moral quality of mind and heart of the moral philosopher who is so sure of who deserves death, castration, and ostracism, and so sure that rational social contracts will provide jobs for executioners, castrators, and deporters" (MP 273). She questions Kant's edict that we should require of ourselves what God would require of us, and to judge ourselves as harshly; and ridicules him for then suggesting that we should cultivate the enjoyment of self-discipline: "This could be seen as deontology's bad conscience, its backhanded concession to a more Epicurean ethics. ... Nietzsche's comment ... seems fair enough: that a bad smell of sado-masochism, the reek of blood and torture, lingers on the categorical imperative" (MP 276-277).

In comparison, Hume is the very model of lenience: According to Baier, his statement that the merits of human beings would be scarcely worth the value of a supper to the righteous and a drubbing to the wicked[9] shows his disinclination to inflict any harm on an agent in retribution for one dereliction, when that harm may cause suffering to the entire character (MP 273-4). On Baier's view, it is not possible to localize punishment within a person so as to reform the vicious part without injuring the virtuous part; and she believes that we should always look "for the social fault behind the individual fault," and take "the responsibility for evil to be shared, never localizable in individual criminals" (MP 288). She does not explicitly draw the implication that we should therefore refrain from punishing individuals at all, but this would be a natural inference.

Baier is particularly incensed by Kant's reasoning regarding how to handle unmarried mothers who commit infanticide. He advocates leniency on the grounds that the victim is not a person whom the law need protect, not on the grounds of sympathy for the mother's social disgrace. While all human deaths require retribution, Kant thinks that God, rather than a human court is the appropriate avenger in this case, since the function of human magistrates is to apply human law to those within its scope.[10] "I should think that Kant's current defenders and admirers must find [this] discussion particularly difficult to recast in a sympathetic way," she comments (MP 277). "This is a pretty shocking and cruel bit of Kantian moral reasoning, cruel in its apparent disregard of the fate of innocent victims" (MP 278).

---

[8] Kant, *Metaphysics of Morals*, trans. Mary J. Gregor (New York: Cambridge University Press, 1991), *Pt. I, The Doctrine of Right*, on the right to punish, pp. 140-145, 168-169.

[9] David Hume, *Essays: Moral Political and Literary*, Ed. Eugene F. Miller (Indianapolis, Ind.: Liberty Classics, 1985), pp. 594-5).

[10] Gregor, *Op. cit.* Note 8, pp. 144-145.

However, the situation only becomes worse, in Baier's opinion, when it comes to Kant's treatment of female unchastity. He regards female chastity and military honor as similar in that both are worthy of defense, and deplores the actual human behavior, and ineffective social institutions, that fall so far short of preserving these ideals in practice. As a counterpoint, she quotes approvingly Hume's explanation, that all human beings, particularly women, are prone to the temptations of pure time preference; and that the real barbarity consists in the hypocrisy and cruelty of men who impose on women the constraint of chastity and the punishment of its violation through shame (T 571-2, cited at MP 278). But since it is too difficult to prove female unchastity or illegitimate paternity in a court of law, shaming a suspect through "bad fame" is the most practical solution (MP 279).

Drawing on Gibbard's analyses of shame and guilt, Baier rightly raises the question of whether shaming a moral derelict is, in fact less cruel as a social sanction than punishing the guilty for a moral dereliction. As Gibbard points out, it is possible to be made to feel shame for things that are not under one's voluntary control, such as one's physical appearance or class background. Moreover, shame is provoked, not by others' anger at one's actions, but by others' disdain and ridicule, which Kant himself describes as "[w]anton faultfinding and mockery, the propensity to expose others to laughter, to make their faults the immediate object of one's amusement, .. a kind of malice, ... in order to deprive [them] of the respect [they] deserv[e], ...[which] has something of fiendish joy in it; and this makes it an even more serious violation of one's duty of respect for other men."[11]

Many psychologists treat anger as itself a secondary reaction to feelings of pain caused by a perceived aggressor, and this links anger and so the blame and attributions of guilt in which it is expressed with perceived harm or wrongdoing. By contrast, disdain or derision expresses merely a disapprobation of someone for failing to live up to the norms of one's group, independently of the moral status or legitimacy of these norms. As Baier observes, "What people feel shame for will depend on what they expect others to sneer or laugh at or treat as grounds for excluding them from some charmed circle of initiates" (MP 279). So shame bears no necessary relation to actual moral dereliction, of the sort that guilt does, and therefore can extend to all of one's perceived flaws, not only the voluntary or moral ones.

The power and undiscriminating sweep of a shame morality leads Baier to ask,

> Who would not opt for Kant's version of the moral world, if the alternative is a social world with some version of a shame morality, where each faulty person faces this threat: somehow get rid of your character fault, or rid us of your faulty presence, or stay and put up with

[11] *Ibid*., Ak. 467.

> our disdain, mockery and avoidance (MP 280) ... Is the Humean morality, which boasts of its nonmoralistic avoidance of 'useless austerities and rigours, suffering and self-denial' ... really so gentle if it condones derision, scorn, disdain, and avoidance of those who do not measure up to to its standards? ... The cost of a nonmoralistic Humean epicurean morality, its may seem, is even higher than that of Kantian moral law enforcement (MP 283).

Baier has six answers to this question in defense of Hume. First, as we have already seen (MP 282), we could not universalize a maxim of social ostracism and ridicule, on pain of incoherence, so shaming practices must be occasional and limited in scope rather than the rule. Second, a Humean can criticize cruel laughter, disdain, and excessive scorn, and the personalities that give them voice (MP 284); and use these shaming practices themselves against their practitioners (MP 286). Third, we can attempt to improve ourselves, not through self-inflicted austerities or self-improvement programs, but rather through changing our circumstances or occupation (MP 284-5). Fourth, we can work to redesign educational and social customs in order to control hurtful derision and criticism of others (MP 284-5). Fifth, we can keep our disdain for others to ourselves, so as not to hurt their feelings (MP 287). And finally, we can, as we have seen, keep in mind to look "for the social fault behind the individual fault," and take "the responsibility for evil to be shared, never localizable in individual criminals;" we can promote "the articulation of shared standards of character assessment, and of application of them to particular persons only when the hurt involved in this application can reasonably be expected to do some compensatory good" (MP 288). In sum, we can limit the damage caused by shaming practices by engaging in them sensitively, infrequently, and wisely; and by reforming educational and social institutions and practices so as to minimize and control them.

4. An Assessment of Baier's Critique

Others besides Baier have criticized the Kantian Contract-Theoretic model for being insensitive to the special and sometimes overriding obligations we may have toward life partners, family, or close friends.[12] But unlike Baier, most have misunderstood their target as metaethical rather than normative. The consequence has often been that they have shot themselves in the foot. They have concluded, self-defeatingly, that we should therefore do away with moral theory, as though they themselves were not doing moral

[12] Notably Lawrence Blum, *Friendship, Altruism and Morality* (London: Routledge and Kegan Paul, 1980); Bernard Williams, "Persons, Character and Morality," in *Moral Luck* (New York: Cambridge, 1981); Michael Stocker, "The Schizophrenia of Modern Ethical Theories," *The Journal of Philosophy LXXIII*, 14 (August 12, 1976), 453-466; Susan Wolf, "Moral Saints," *op.cit.* Note 5.

theorizing in claiming this; or with impartiality, as though they did not intend these very claims to apply impartially; or with moral ideals, as though the compassionate acceptance of agents' moral flaws and imperfections were not itself a moral ideal. None have suggested, as Baier does, that a familial model of interconnected but unequal power relationships among morally imperfect agents itself should be the measure of moral analysis.

This is a much harder suggestion to refute, because it conforms so much better than the Contract-Theoretic model, even in the professional and market spheres, to the actual facts of our lives. Baier stresses the neglect by the Social Contract-Theoretic tradition of the roles of women and children. But in reality it applies no more closely to the roles of actual men, who are just as dependent, unequal in power, and involved in intimate relationships as women. That Baier's Humean familial model is so much more directly relevant does not make it any easier to live up to – to refrain from emotionally abusing our children when they get on our nerves, for example; or from publicly one-upping our spouse for their household incompetence out of barely repressed hostility; or from undermining a colleague who is treading on our professional turf. So it does not quiet the typical Anti-Rationalist complaint that normative moral theory is too demanding, nor license its conclusion that we should just all relax and do what we like. Baier's recommendations are just as demanding, if not more so; and they are very far from licensing us to relax and do what we like. The challenge Baier's thesis presents is rather to show why the Social Contract-Theoretic model should be retained at all, if it bears no realistic relationship to what we actually are and do.

Kant's favorite tactic for de-fanging Hume is the one I am deploying throughout this project: to accept but subsume Hume's analysis of something – causality, substance, the self, practical reason – as a special case under one that is deeper and more comprehensive in scope. As a good Kantian, I believe this tactic will work particularly well against Baier's Humean analysis. Specifically, I think it can be shown that there is plenty of room for both the familial and the Contract-Theoretic model within a fully elaborated moral theory; and that the relation of the familial model to the Contract-Theoretic model is one of lower-level generalization to higher-level theoretical construct within such a theory. I do not try to defend this claim until Volume II, Chapter V.5.2. But I can try here merely to suggest some ways in which Baier's recommendations for revising the central orientation of normative moral theory are not as incompatible with Kantian Social Contract Theory as she makes them out to be.

The importance for normative moral theory of the question of how to raise children to be responsible moral agents and so insure the moral stability and continuity of a community, and of how this essential activity should be conceptualized within any such theory that purports to have practical

application, cannot be overestimated. Baier is right to bring this issue to the foreground. Even if we knew what perfect parenting would look like, which we don't, we still would have no way of bootstrapping ourselves into that ideal, and so of avoiding the cross-generational transmission of moral defects and dysfunctions to our children. Under these circumstances, how should we conceptualize our own moral roles as concerned parents, teachers, and citizens? And what, then, is the point of normative moral theory, other than ineffectual wishful thinking?

Now Baier argues that the primacy to moral theory of child-rearing implies that obligation cannot be the primary concept, since we cannot make sense of an obligation to love our children. But this does not follow. For surely the one question we must all ask ourselves in response to Baier's challenge is what, given that none of us are paragons of love, patience, or sympathy, our obligations to our children realistically are. And it is hard to avoid the Kantian answer that at least part of our obligation is to cultivate these virtues in ourselves to the best of our abilities, as well as to work to reform social and educational institutions that can make these obligations easier to discharge.

Among these obligations, of course, would be that of raising our children to take promise-keeping seriously, and so to become "morally competent promisors." But that kept promises are compatible with more severe moral betrayals does not show that keeping promises is not an essential and basic moral obligation. One difference between a morally competent promisor and a moral hypocrite is that the former does not adhere to the letter of the law while violating its spirit: does not, for example, fulfill his contract to fix your plumbing while blowing up your house, or tune up your car while destroying its muffler, or promise to write you a favorable job recommendation that in fact damns you with faint praise, or to mediate your domestic conflict while seducing your spouse. Unlike a moral hypocrite, a morally competent promisor does not cross her fingers behind her back, or silently recite escape clauses to herself under her breath while she is making you a promise.

Particular promises have meaning and stability only against a background of shared assumptions about what constitutes honest and reliable behavior. Morally competent promisors share the assumption that making and keeping a particular promise is an instance of an effective, pervasive, and rule-governed social practice of promising, not an exception to a pervasive social practice of dishonesty, betrayal and opportunism. To show that this assumption may be misplaced under certain circumstances does not show that promise-keeping is not a central or definitive moral obligation, nor that obligation is not the primary concern of imperfect moral agents who are obligated to do their imperfect best.

Baier is also to be commended for highlighting the practical moral importance of the intimate, dependent, involuntary and unequal power relationships in which we are embedded – as parents, children, spouses,

colleagues, friends, and citizens. And she is surely right to argue that analysis of the use and abuse of power and responsibility should play a much more prominent role in the construction of normative moral theory than it does. Yet does this imply the irrelevance, or even the noncentrality to moral discourse, of the Contract-Theoretic concepts of freedom, autonomy, individuality, equality and individual rights? I am not convinced. For it is only with the aid of these concepts that we can understand what is wrong with neglecting the important contributions of women and mothers to the creation of a moral community in which the benefits of freedom of choice, voluntary association, or autonomy of life-plan are available to others; or in which these benefits might be rejected for others in which care, concerned intervention in the lives of the powerless, and grass-roots political work in one's community of origin rather than one's community of choice have greater priority.

Similarly, it is only with the aid of the Social Contract-Theoretic concepts of freedom, equality, individuality, autonomy and rights that we can understand what is wrong with failing to respect the special needs of children, or with failing to treat them as individuals, or with assigning them too much or too little power and responsibility for themselves and their environment in relation to their age. And it is only with the aid of these concepts that we can come to understand the responsibilities and obligations of those with greater power – in the home, workplace, marketplace, or social or civil sphere – to those who have less. It is hard to see how we could come to grasp the moral implications of unequal power relations in any of these contexts, without these background Contract-Theoretic concepts to give them meaning.

What we see from Baier's analysis is that women, mothers, children and other disenfranchised groups have the same rights as those who traditionally have arrogated those rights to themselves. But we need the background Contract-Theoretic concepts of rights, freedom, equality, and individuality as a measure in order to understand what is amiss. Of course this does not imply that everyone in actual fact is or should be exactly alike in degrees of freedom, autonomy, or power. No Social Contract Theorist claims this. That traditional Social Contract-Theoretic assumptions do not apply concretely and realistically to actual human agents does not show that they are not central to moral theorizing. What it shows is that they cannot be expected to do the practical work Baier rightly expects the lower-level generalizations of a normative moral theory to do. But no Social Contract Theorist claims that they can.

Now we have also seen in Section 2 that Baier has particularly harsh words for Kantian Rationalism, i.e. the assumption that all we need do morally is to subject our passions, whatever they are, to rational control. Kant himself did not make this claim, and I know of no Kantian who has. Kant himself devoted a great deal of thought to the question of how to cultivate

those moral virtues of feeling and sensibility that would make effortless the enactment of duty,[13] and in fact presupposed love of others, and sympathy, as a precondition of subjection to the moral law.[14]

But again it is hard to see how Baier's agenda of replacing rational control with parental love as a condition of raising morally healthy agents can ever get off the ground, if we are not even allowed to exercise rational control in order to cultivate the parental love she advocates. We are none of us angels of compassion, and often do feel – in addition to love – irritation or anger at our children. We can either vent that anger at our children, or we can control it. Baier cannot possibly think it would serve the cause of raising morally healthy agents (or moral authenticity, either) to make a practice of venting it. So what would she have morally committed parents and teachers do? Evacuate their children to an air raid shelter when they feel an outburst of temper coming on? Give them up for adoption if one judges oneself too irascible to raise them?

What she says we should all do, under such circumstances, is to exercise Humean passional reflection rather than Kantian rational control. The problem here is that her analysis of Humean reflection is not sufficiently distinguishable from Kant's actual conception of reason to do the job. Take her description of reflection as "a response to a response ... a sentiment directed on sentiments" (MP 72), and Hume's own characterization of passions as "'returns upon the soul' of remembered experience of good and of evil" (MP 83). We can feel a desire for a desire, or anger about our anger, or gladness for our joy. But to desire a desire is to desire an intentional object we must be able to identify as a desire. To feel angry at our anger is to react to an intentional object we must be able to identify as our anger. In the first case the object of our intentional attitude may or may not be a remembered experience; in the second case it is. In either case, in order to remember the experience of good or evil that now returns upon the soul as the object of our desire or anger respectively, we must be able to conceive it as being the kind of object it is. This point merely generalizes to all intentional objects the representational analysis of desire offered in Chapter II.2.1. In all such cases, we must be able to judge the intentional object by ascribing predicates to it as subject in a categorical indicative judgment.

This is the essence of reasoning for Kant. It is not necessarily linguistic, but it is necessarily conceptual; i.e. it involves what Hume would call "ideas." And in order to have had the experience that now returns upon the soul in the first place, and to which we are now responding passionally, we had to have made a similar judgment about the extrinsic state of affairs to which that

---

[13] *Op. cit*. Note 8, *The Doctrine of Virtue*, Ak. 387-388; Ak 391-3; 457; 477-486.
[14] *Ibid*. Ak. 399, 401-2, 456-8.

experience itself was a response. So Kantian judgment, and therefore reason, is intrinsically involved in our passional responses, both to extrinsic states of affairs and to other passions themselves.

Hume would not deny this. Baier seconds Hume's "excuse" for beginning his analysis with ideas rather than passions, which is that most passions depend on ideas, and "as it is by means of thought only that any thing operates upon our passions, and as these are the only ties of our thoughts, they are really to *us* the cement of the universe, and all the operations of the mind must, in a great measure, depend on them" (T 662). Baier interprets this passage as showing the primacy of practical reason that theoretical reason serves (MP 77). But this is not what Hume says here, nor what the passage implies. What Hume says is that passions presuppose the ideas to which they are responses, and that the order and association of ideas presuppose the passional responses that connect them. Passions respond to ideas, and ideas are associated by passions. What the passage implies, then, is that ideas and passions are mutually dependent. But if passions respond to or are caused by ideas, and actions respond to or are caused by passions, then by transitivity of causality actions respond to or are caused by ideas. So not only does this passage not claim the primacy of practical over theoretical reason. It implies just the opposite.

Now Kant would not disagree that, so far as empirical experience is concerned, ideas, i.e. particular categorical indicative judgments, do, indeed, seem to be associated by passions or other contingencies. These are the empirical mental habits and customs to which Hume reduces reasoning. Kant would simply add that a necessary precondition for such empirical mental experience is the *a priori* connection of these ideas by the transcendental forms of judgment he lists in the Table of Judgment in the first *Critique*.[15] For Kant, these forms of judgment describe the ways in which genuine reasoning must occur; that is what makes them transcendental. And Kant tells us that reason in general, not just the categorical imperative, compels our respectful attention (*Achtung*), if not always our submission to its dictates.

Baier's account of reflection, and particularly her description of our attempts to ascertain whether those mental habits and customs to which she and Hume reduce reason can "bear their own survey," illustrates the subordination of empirical mental activity to the transcendentally rational constraints on which Kant insists. As we have seen in Section 3, Baier describes this survey as itself administered by the passions, including socially dependent ones (MP 82). And she argues that we can do no more than repeat this operation at ever higher levels (MP 84). But if this were all there were to it, there would be no point in conducting the survey in the first place, since we

[15] *Kritik der Reinen Vernunft*, herausg. Raymund Schmidt (Hamburg: Felix Meiner Verlag, 1976), A 70/B 95.

could always ask the same questions about the passions administering the survey as about those being surveyed. As we have already seen in discussing Frankfurt, we would be stuck in an infinite regress.

So when we feel compelled to "[think] long and hard about [a mental habit's] nature, its sources, its costs and its consequences" (MP 81), and to ask ourselves the question, "[H]ow important do we on reflection judge it that [people] have [this habit of mind]?" (MP 82), we are not evaluating our associative mental habits against the criteria of other associative mental habits that are just as contingent and empirical. If we were, the questions just quoted would not have the special urgency, status or finality in such a survey that they do. Questions that invoke the nature, causes, costs, consequences and importance of a thing by definition call on rational criteria for evaluating it. Other questions we could ask, for example, about a mental habit's frequency, content, duration, and simplicity, or about how amusing we on reflection judge it that people have this mental habit, do not.

Clearly these latter questions are irrelevant to Baier's reflective survey, even though they may just as passionally associated. But any attempt to analyze the passional association of ideas that necessarily is *not* irrelevant to serious reflective survey will invoke precisely the rationality criteria that Baier claims to have jettisoned. So what we are doing, rather, is heeding the demand of reason – in Kant's transcendental sense – that all such associative empirical mental habits conform to its constraints. We are evaluating our empirical mental habits with reference to criteria that recognizably constitute the concept of rationality. So just as for Kant, Hume's passions, and the passional association of ideas, must bear the survey, not of other passions, but of reason; and Baier does not provide a genuine alternative to the rational control of them she claims to reject.

Is the Kantian requirement of rational control, and its background normative theory, more cruel than Hume's "morality with a light touch"? In the end, Baier does not seem to think so. She expresses great indignation for Kant's support of *Lex Talionis*, but then faults him for his unwillingness to apply similarly harsh justice to unmarried mothers convicted of infanticide. She castigates punishment for its contagiously harmful effects on already faulty character, and for erroneously individualizing moral fault that in fact is shared. But she advocates it anyway, on cost-benefit grounds. And after detailing, with great honesty, the hurtfulness of the shame-based morality to which she believes Hume to be committed, her six defenses of Hume merely describe ways in which the damaging effects of such a morality might be contained. There is no suggestion that this morality is any less cruel than it might first appear, and none that it is, in the end, less cruel than Kant's.

In fact, in the end she concedes the role of punishing moral derelicts to Kantian magistrates when she says that "[t]he deliberate, contrived, and calculated hurting of the faulty person, in order to reform him, will be left to

magistrates and other licensed coercers. The good Humean will want to keep a careful check on how such persons (parents, schoolteachers, judges, police, prison guards) are exercising this grave responsibility. Morality must be solemn when it ... endor[ses] society's self-protective coercive activities" (MP 288). So it seems that Baier's Humean morality ascribes much greater value to rationality, justice, "stern imperatives", and just punishment than first appears. To find a genuine alternative to the Kantian Social Contract Theory she deplores, we must turn to Baier's substantive analysis of trust.

5. Baier's Analysis of Trust

Appropriate trust, as Baier characterizes it, is centrally connected with the notions of good will, power, and making oneself vulnerable to another. She defines trust as a belief-informed and action-influencing attitude (MP 10, 132) that makes one more vulnerable to harm from another, in the confidence that the other will not use their discretionary power to harm one because there is no reason to do so (MP 11, 133, 152, 187 fn. 9); as reliance on another's good will; and as a vulnerability one accepts to another's possible but not expected ill will (MP 99, 105). Relying on someone is nevertheless different from trusting them, in that the former concerns only their dependable habits, whereas the latter concerns their good will (MP 98). Trust involves allowing others to exercise discretionary powers to take care of something (or someone) one cares about (MP 105). Trust involves the paradox that "in trusting we are always giving up security to get greater security, exposing our throats so that others become accustomed to not biting" (MP 15).

Trust becomes pathological when, first, the enterprise whose workings trust improves is evil, as are the trustworthy members of a death squad (MP 131). Second, the enterprise in general may be benign, but its treatment of some of its members unfair. In that case, their trust is equally unhealthy, as is the trust employees may feel toward an employer "whose exploitation of workers is sugar-coated by a paternalistic show of concern for them and the maintenance of a cozy familiar atmosphere of mutual trust" (MP 131). Third, the attitude of trust can be faked, and backed instead by vigilance or threat advantage, as is a wealthy wife's who suspects her husband of adultery (MP 132).

Trust may become unhealthy, fourth, when the trustor is too quick to call the trusted to account; and fifth, when the trusted misuses her discretionary powers, perhaps through laxity or risk-taking, or, at the other extreme, through reliance on a rigid rule that excludes discretionary power altogether, as when one's spouse can be trusted to remember one's birthday because he has given his bank a standing order to send flowers every year on that date (MP 136). Sixth, trust may degenerate into mutual predictability, when what should have depended on discerning judgment as to the trustor's best interests becomes merely a reflexive habit of behavior, as when a university

administration distributes equal benefits to every department because it would involve unwanted work to find out which ones deserved them and which ones did not (MP 136-138). Seventh, trust may dwindle into constant vigilance, checking and testing of the trusted's capacity to carry out her responsibilities; or, eighth, may balloon out of proportion for fear of insulting the trusted by requiring any checking or testing whatsoever (MP 139). Pathologies of trust and distrust, Baier concludes, "occur where there is the will to monopolize and hang on to power, to keep the underdogs under, to prevent inferiors from advancing" (MP 147).

One way of violating trust is by taking on the care of more than one was entrusted with (MP 101), as does, for example, the secretary who makes it his business not only to retype one's manuscript, but to edit it for felicitous phrasing. So trusting someone also involves trusting them to recognize the limits of the discretion entrusted to them; and the more discretion they have, the harder it is to determine when those limits have been exceeded (MP 103). Trust can also be violated, not only through ill will, but also through incompetence or negligence, as for example, happens when you entrust a friend with a confidence who then forgets its confidentiality and relates it as an entertaining anecdote at a party. Concealing ill will under the cover of discretionary use of trust, or of incompetence, is yet a further violation (MP 104-5, 135).

What is the difference between morally justified trust and foolish trust? When is my trust in someone warranted, and when is it properly undermined? Trust is rational, for Baier, if there is no reason to suspect in the trusted overriding motives that conflict with the demands of trustworthiness as the trustor sees them (MP 121). So, for example, if a husband has reason to suspect that his wife is raising their daughters to dislike men, he has reason to withdraw trust in his wife's decisions; or if her motives for not doing so are no longer outweighed by the anticipated costs to her of his withdrawing his economic support. The husband must judge whether and how ambivalent his wife's motives are in order to establish whether it is rational to continue to trust her (MP 122). Rational trust is compatible with some degree of suspicion or vigilance. But when the trustor must rely solely on threats or pressure he can exert to maintain the relationship, or when the trusted must rely on concealing breaches of trust, then the relationship is morally rotten, and can be expected to deteriorate when these facts are made explicit (MP 123). So a trust relationship is morally decent, Baier suggests, when no such threats or concealments are necessary (MP 123, 124, 128). If the relationship can survive the mutual awareness by both trustor and trusted of each other's reasons for being able to rely on the other to continue the relationship, then the relationship is morally sound (MP 128).

But in the end, there can be no rules about when or where or whom to trust; to what extent we should rely on or question our instincts about those

we instinctively distrust; or how often we are justified in forgiving a betrayal of trust (MP 139, 141-142). And rigid rules and algorithms will not help wise individuals or a wise society guard against the untrustworthy (MP 160). Nevertheless, giving another the benefit of the doubt as a rule of thumb is justified on the grounds that "[t]here are few fates worse than sustained self-protective self-paralyzing generalized distrust of one's human environment. The worst pathology of trust is a life-poisoning reaction to any betrayal of trust" (MP 145), given that "the trust-dependent goods are the most precious" (MP 146).

Moreover, to protect our ability to trust we can use our powers of judgment, both in case-by-case decisions and in the design and overhaul of social institutions (MP 160, 176), and the rules and recipes we have derived from experience for designing lasting schemes of cooperation (MP 161). "'Ingredients' such as empowerment of the more vulnerable, equal respect, balance of power, provision for amendment, a place for the hearing of grievances, all give us ideas that we could try incorporating into rules for the design of other stable schemes of trust-involving cooperation, so that all trust comes closer to being mutual trust, and so also to mutual vulnerability" (MP 161-2).

There are also some forms of trust that strengthen and extend the practice of trusting, for example, those in which trust is reciprocal and both parties are at least roughly equal in power and vulnerability. Contract, solemn vows, the appointment of a godparent, guardian or trustee are ways in which the more powerful can selectively disempower themselves and so avoid the temptation to abuse or manipulate their power (MP 178). Trust in one's own judgments of trustworthiness, what Baier calls "meta-trust," is almost always, she thinks, preferable to the ultimately disabling effects of distrust (MP 185), even though "[t]he more one knows about people (oneself included), the less one has occasion strictly to trust them, or to trust trusting them" (MP 187). Finally, trust in sustained trust, "in full knowledge of its risks as well as its benefits, and trustworthiness to sustain trust may well be the supreme virtues for ones like us, in our condition" (MP 185, 197, 201). In the end, Baier thinks, we simply have to make a commitment to value and practice it, despite the risks and betrayals, in order to further a community grounded in trust, in which such risks and betrayals are minimized.

On Baier's view, the concept of trust brings together "men's theories of obligation" with "women's hypothetical theories" of love and care (MP 10), by generalizing central moral features of obligations, virtues, and loving along with such relationships as those between teacher and pupil, confider and confidant, worker and co-worker, and profession and client (MP 15). A moral theory that spelled out the conditions for appropriate trust would, therefore, include a morality of love, as well as supplement a morality of obligation: "to recognize a set of obligations is to trust some group of persons to instill them,

to demand that they be met, possibly to levy sanctions if they are not, and this is to trust persons with very significant coercive power over others. ... the morality of obligation, in as far as it reduces to the morality of coercion, is covered by the morality of proper trust." What obligations we have and what virtues we should cultivate is a matter of what it is reasonable to trust ourselves to demand and expect from one another (MP 12), so not only obligations but virtues as well presuppose the concept of trust. Indeed, virtually all aspects of implementing a moral code within a community, on Baier's view, depend whom we can trust, and how far, to do what jobs, whether parental, political, coercive, or social (MP 14).

One reason why the concept of trust is so foundational to moral theory, on Baier's view, is because it is so pervasive. Threats and coercion could not be the main support for a moral code, she argues, for fear of an infinite regress: Either there must always be more coercers to threaten the coercers to do their job of coercing our compliance, or else we must finally trust some such coercers to do so without any such backups (MP 14, 164). Thus trust conditions our relationships not only with friends and family and colleagues, but with strangers and even with enemies (MP 98). In all such cases, we observe conventions of behavior that we trust others not to violate – for example, not to give us false directions when we ask for help in a foreign city, and not to shoot after we have waved the white flag in war. "We trust those we encounter in lonely library stacks to be searching for books, not victims. We sometimes let ourselves fall asleep on trains or planes, trusting neighboring strangers not to take advantage of our defenselessness. We put our bodily safety into the hands of pilots, drivers, and doctors with scarcely any sense of recklessness" (MP 98). In such cases we trust others not to violate our persons, our property, or our autonomy (MP 103).

Contract-Theoretic morality traditionally has focused on the very limited form of trust involved in promising and the fulfillment of contract, Baier thinks, because most of the great moral theorists – she cites Hobbes, Butler, Bentham, and Kant – were "a collection of clerics, misogynists, and puritan bachelors ... who had minimal adult dealings with women," and so made central to their moral analyses "cool, distanced relations between more or less free and equal adult strangers, say, the members of an all-male club, with membership rules and rules for dealing with rule breakers and where the form of cooperation was restricted to ensuring that each member could read his *Times* in peace and have no one step on his gouty toes" (MP 114).

By contrast, those who are more engaged socially as lovers, husbands, and fathers with women, the ill, the very young, and the elderly – here she cites Hume, Hegel, Mill, Sidgwick, and maybe Bradley – will have a more complex view of moral relations (MP 114). The reason the great moral theorists did not focus on the forms of trust that embed one in dependent, unequal, noncontractual and nonvoluntary social relationships was that they

were too invested in collectively exploiting women in those roles to confront their own moral hypocrisy (MP 115). The prevailing interest in current moral philosophy in Prisoner's Dilemma problems carries the obsession with "moral relations between minimally trusting, minimally trustworthy adults who are equally powerful" into the contemporary context (MP 119).

But Baier argues that Social Contract-Theoretic morality cannot explain trust when the relation between trustor and trusted is radically unequal in power, as between infant and parent; and that it is the negligence of such relations that has led philosophers to think that trust can be explained in Social Contract-Theoretic terms (MP 106, 109). "Contract," she maintains, "is a device for traders, entrepreneurs, and capitalists, not for children, servants, indentured wives, and slaves. They were the traded, not the traders, and any participation they had in the promising game was mere play" (MP 113). She applauds Nietzsche for stating plainly what other liberal theorists refused to acknowledge, i.e. that Social Contract Theory must conceive women in their traditional roles as housewives and mothers as property and as predestined for service. And even when women are fully equal to men as moral subjects, the continued expectation that they will take responsibility for bearing and caring for children displaces the central role of voluntary agreement to Contract-Theoretic morality. "Since a liberal morality both *must* let this responsibility rest with women, and yet cannot conceive of it as self-assumed, then the centrality of voluntary agreement to the liberal and contractarian morality must be challenged once women are treat as full moral fellows" (MP 113-114).

By contrast with the Contract-Theoretic conception of trust, infant trust is comparable to trust in God, for Baier, in that it is total and fully dependent. It is unique in that it "normally does not need to be won but is there unless and until it is destroyed .... Trust is much easier to maintain than it is to get started, and is never hard to destroy. Unless some form of it were innate, and unless that form could pave the way for new forms, it would appear a miracle that trust ever occurs" (MP 107, 195). As Baier observes, this fact makes it difficult to see how the hypothetical Hobbesian conversion from distrust in the state of nature to mutual trust in the social covenant can be psychologically realistic. Once a child realizes that her parents are not God, the best reason for her to continue trusting them to take care of her best interests is that she sees her best interests as a good for them, too; i.e. that she feels loved (MP 108). As she approaches adulthood and her parents approach dependent old age, the power relations between them may equalize and then imbalance in the other direction. But at no stage, Baier maintains, can these relations, even at their most equal, be understood in terms of a mutual contractual exchange (MP 109).

Moreover, any account of trust that makes that between infant and parent central cannot depend on the use of concepts and abilities that an infant does

not have (MP 110). This means that such an account cannot project or "reconstruct" trust in terms of voluntary acts, consciously acknowledged risks, or deliberately imposed controls on the damage done by potential violations of trust (MP 110-111). Rather, the initial assumption must be of automatic or unconscious or unchosen mutual trust. And "[w]hat will need explanation will be the ceasings to trust, the transfers of trust, the restriction or enlargements in the fields of what is trusted, when, and to whom, rather than any abrupt switches from distrust to trust" (MP 111). This is why Baier's analysis of trust is silent on matters of choice and deliberation, and why she doubts the ability of Social Contract Theory, which relies on such assumptions, to explain it. "[T]rust," she observes, "in those who have given us promises is a complex and sophisticated moral achievement" that takes for granted less artificial and less voluntary forms of trust, for example, in friends and family, and in others sufficient to engage with confidence at least in minimal simultaneous transactions and exchanges (MP 112).

6. An Assessment of Baier's Analysis of Trust

In this brief summary I have not done justice to the subtlety and scope of Baier's treatment of the topic of trust, nor to the range of significant and complex moral phenomena she brings to our attention. Baier maintains that a women's (normative) moral theory, of which she thinks there are none, "will need not to ignore the partial truths of previous theories. It must therefore accommodate both the insights men have more easily than women and those women have more easily than men. It should swallow up previous theories" (MP 4). In her vehement rejection of Kantian Social Contract Theory, it cannot be said that her analysis attempts to meet these criteria, though I have already suggested that it might to a greater extent than she would prefer.

On the other hand, Baier's analysis does seem to satisfy her description of "paradigm examples of moral theories." These, she claims, have a "broad brushstroke" comprehensiveness and coherence, a fairly tight systematic account of a large area of morality, with a keystone supporting all the rest," which are, she thinks, the antithesis of the "mosaic method" of "assembling a lot of smaller-scale works until one [has] built up a complete account" (MP 3). This distinction certainly will not work to classify Rawls, our quintessential moral theorist, in the right way, since his theory of justice has both sets of features, the first with regard to structure, the second with regard to temporal process of construction.[16]

[16]The early "building blocks" of Rawls's *Theory of Justice* (Cambridge, Mass.: Harvard University Press, 1971) include "Outline of a Decision Procedure for Ethics," *The Philosophical Review 56* (1951), 177-197; "Justice as Fairness," *The Philosophical Review 57* (1958); "The Sense of Justice," *The Philosophical Review 62* (1963); "Constitutional Liberty and the Concept of Justice," *Nomos VI: Justice*, Ed. C. J. Friedrich and John Chapman

So does Baier's. Her approach is mosaic, through the series of essays in which *Moral Prejudices* consists. It is true that her account is less tight and systematic than the above summaries would make it seem. As is clear, I have, in this exposition, extrapolated claims and arguments from a large variety of her essays and reorganized them according to my own requirements of system. Nevertheless, the elements of coherence and comprehensiveness of the area of morality she covers, and even the "keystone supporting all the rest" are there for any reader to see. The keystone is, of course, her analysis of trust, and the moral norms it generates by way of her analyses of pathological trust, betrayals of trust, rational trust and morally decent trust. There are even what I described in Chapter V.1.2 as practical decision-making principles in her account of how to protect our ability to trust.

Moreover, Baier makes a forceful case for the comprehensiveness of her account by arguing, first, that the concept of trust is presupposed by those of virtues and obligations; second, that its pervasiveness trumps Contract-Theoretic attempts to ground social equilibrium in coercing the fulfillment of obligation, by capping a threatened infinite regress of coercers; and third, that it can explain important moral phenomena, namely moral transactions among agents unequal in power, that Social Contract Theory cannot explain. One does not have to agree with these arguments to see that she has made the case.

Baier offers further criteria that a genuine moral theory of trust should meet when she claims that "[a] moral theory which made proper trust its central concern could have its own categorical imperative, could replace obedience to self-made laws and freely chosen restraints on freedom with security-increasing sacrifice of security, distrust in the promoters of a climate of distrust, and so on" (MP 15). These are criteria her own account of trust does, in fact, satisfy. Its categorical imperative may be identified as "trust in sustained trust and trustworthiness to sustain it" (MP 185, 187).

Now Baier withdraws this suggestion on the next page with the disclaimer that she is "not really concerned to elevate any virtue to supremacy. Even if we could effect some sort of unification of the virtues by relating them all to due trust and due trustworthiness" – and she then proceeds to point out several of the frequently neglected virtues such as tact, discretion, resilience, and alertness to the oppression of the silenced, that would become more salient relative to trust relationships – "we will still need a whole host of virtues, more or less democratically ruling in our souls,

---

(New York: Atherton Press, 1963); "Distributive Justice," in *Philosophy, Politics and Society*, Third Series, Ed. Peter Laslett and W.G. Runciman (Oxford: Basil Blackwell, 1967); "Distributive Justice: Some Addenda," *Natural Law Forum 13* (1968); and "The Justification of Civil Disobedience," in *Civil Disobedience*, Ed. H. A. Bedau (New York: Pegasus, 1969).

balancing each other's likely excesses. ... it is not part of my present aim to show that due trust and due trustworthiness can lord it over other virtues. My aim here is less imperial, more modest – to imitate Hume ..." (MP 188).

But in this instance Baier's characteristic humility calls Uriah Heep too vividly to mind. Her call for the "democratic rule" of our souls by "a whole host of virtues, ... balancing each other's likely excesses," cannot be given too much weight without a solution to the problem of moral paralysis such a motivational cacophony would generate. Similarly, her worries about the "imperialism" of "elevating any virtue to supremacy" cannot be taken seriously, first of all because there is nothing wrong with the so-called "imperialism" of trying to construct a unified and comprehensive moral theory,[17] and second of all, because Baier has done precisely that. Whether she likes it or not, and whether one agrees with it or not, she has produced a genuine moral theory according to her own criteria of what one would look like. And if she really thinks no woman before her has done so, then she is the first. She might just as well stop squirming and get used to her achievement.

There, is of course, more to say. There are certain fundamental cases of trust relationship that Baier does not discuss; and to bring these up is to again raise the question of whether Baier has succeeded in distancing herself from the Kantian Contract-Theoretic model as fully as she wishes. In particular, Baier does not discuss the vulnerability a moral wrongdoer displays to her victim, and the trust in her victim's moral rectitude the wrongdoer must have, in order to admit guilt, apologize, or request forgiveness. The ability to perform each of these acts presupposes that the victim can be trusted to show compassion and refrain from using the wrongdoer's vulnerability to shame, punish, or exploit her. And to presuppose that the victim can be trusted in these ways is no more or less than to presuppose that the victim will continue to fulfill certain moral obligations toward the wrongdoer despite the moral wrong done him. That is, it is to presuppose that the wrongdoer does not, in the very act of harming the victim, drag the victim down to the wrongdoer's level. What the wrongdoer needs to trust in order to thus make amends, is that the victim will retain the role of moral agent even after the wrongdoer has deserted it; that the victim will not be, literally, demoralized by the wrong done him.

The prevalence of moral wrongdoing, writ small as well as large, makes this type of moral vulnerability exceedingly widespread. We each experience it in the defensiveness with which we react to the mere suggestion that we have acted unethically in any respect, as though it were realistic to conceive

[17] There has got to be a limit to how much political correctness even the most politically interested among us can stand. Can Baier in fact really believe, not only that no women *have* constructed a moral theory (MP 2), but that, because of its Kantian approach and "imperial" ambitions, no women *should*?

ourselves as paragons of moral virtue in the first place. But the distrust of putting oneself further in one's victim's power by owning responsibility and apologizing for one's wrongdoing – and thereby making oneself even more vulnerable – also can be measured by the extreme psychological difficulty of doing so. This is evidence of the extent to which a shame morality in fact prevails. It is just too frightening and dangerous to put oneself at risk with another person to that extent, and too traumatizing even to consider submitting oneself to the shaming and ego-crushing punishment the other could then inflict. Consequently, such behavior is very, very rare. Certainly no one could plausibly argue that owning responsibility and apologizing for moral wrongdoing is, in this culture, a prevalent and motivationally effective norm. The humility, trust, and moral spine needed to perform these acts, and the moral tolerance by the victim they presuppose, demand far too much of most ordinary moral agents.

Far more common, when one comes to recognize one's moral wrongdoing, is simply to then treat the victim as though no such wrong had been done; and dismiss the victim as irrational, or punish him further, if he protests or refuses to collude in this fantasy. The purpose of the pretense is to impose forgetting as a substitute for forgiving, and thereby to fortify one's delusion of moral impregnability against possible attack. This makes any reproach or blame from the victim appear as an undeserved blot on the wrongdoer's unblemished moral character. Creating and maintaining the delusion of moral invulnerability through pretense and willed amnesia is a well-established custom in European-American culture for deflecting responsibility, not only for personal but also for institutional and historical wrongdoing against victimized groups and individuals.

For example, here is an anecdote to rival even the goriest Baier serves up. An African American woman writer was invited by a group of European American colleagues to contribute an essay to an edited collection. She accepted, and sent them a summary of the topic on which she proposed to write. She received in response a letter from the least professionally powerful of the editors, accepting and applauding her proposal. Shortly thereafter she received another letter from all of them requesting that she write on a different topic because, in their opinion, she would not be able adequately to defend her chosen thesis. When she wrote back to protest the arrogance and condescension of this judgment, she received a third letter from all of them, explaining that the real problem was that her chosen topic was too controversial and against the grain of received interpretation. After she wrote again to protest this outright attempt to suppress her work, she learned from the publisher of the volume that her name and abstract had been dropped before the book proposal was sent to outside referees. When she again wrote to protest that she was not being treated fairly, she received a hostile and menacing letter from the most professionally powerful of the editors, advising

her that if she wanted to continue to participate on any level in this project, she was to drop the matter immediately and say no more about it. After repeated and vehement written protests against this behavior, which by now included not only arrogance, condescension, and censorship but also verbal abuse and intimidation, she was finally sent a contract by the publisher. The next time she encountered the editors at a professional function, they greeted her with a smile and a wave. One remarked how nice it was to see her.

Now of course there are many reasons why the editors might have declined at any point simply to admit wrongdoing and apologize for their behavior. One possibility, of course, is that they did not regard her as worth apologizing to, any more than I would consider apologizing to an ant I had squashed while walking through the grass. Another is that they did not at any point regard themselves as having done anything to apologize for, any more than I would think I had committed a moral wrong by squashing the ant in the first place. But the most plausible explanation is that, having already made themselves morally vulnerable to her by revealing their unpleasant beliefs and motives, they regarded it as too risky to increase their moral vulnerability to someone they now expected to use it in retaliation. So in defense of that anticipated retaliation, they chose instead to compound their moral dereliction with hypocrisy and evasion of accountability. Thus, their moral wrongdoing – itself an expression of mistrust and uncollegiality – made it impossible for them to trust her not to behave as contemptibly toward them as they had toward her. Their moral wrongdoing itself destroyed any possible climate of trust in which an apology would have been a realistic option, and exacerbated the extent of the wrong they committed.

This is the kind of case that illuminates the sense in which trust always depends on the fulfillment of reciprocal *obligations* between trustor and trusted. A necessary condition of the trustor's extending trust to the trusted is that the trustor not consciously have offended morally against the trusted in such a way as to undermine the trustor's own belief in the trusted's essential good will toward him. That is, the trustor is obligated to treat the trusted with respect as an end in herself. Now it may happen that the trusted may be of such a psychological makeup that no amount of respectful treatment is sufficient to guard against the development of suspicion and ill will toward the trustor. I discuss this possibility at greater length in Volume II, Chapter XI.4. But this contingency does not vitiate the trustor's obligation to do the best he can.

Similarly, a necessary condition of being genuinely trustworthy, as opposed to merely seeming that way, is that the trusted be prepared to act in such a way as to continue to be worthy of such respect, even if it is not forthcoming. That is, the trusted is obligated to uphold those "stern imperatives" against moral wrongdoing even if she is the victim of it. Again it may happen that the moral wrongdoing the trusted suffers is so insulting,

painful or damaging that it simply is not in her power to continue to conduct herself with moral equipoise in its aftermath, rather than lash out in self-defense against those who have inflicted it. But again this does not relieve her of the obligation to do the best she can.

These obligations hold regardless of the relative power imbalances between trustor and trusted. Infant trust trivially satisfies these requirements, regardless of how counterintuitive it may be to conceive of infants as having obligations. If, in accordance with Baier's strictures against accounts of trust that invoke concepts and abilities infants do not have, infants cannot have obligations because they cannot be said to act, they also cannot morally offend against the trusted for the same reason. There is nothing infants can do to their caretakers that would undermine their instinctive (but nevertheless rational) belief in their caretakers' essential good will toward them. This is part of why an infant's unconditional dependence can inspire a caretaker's unconditional love.

On the other side, a necessary condition of being a good parent or caretaker of infants is that one continue to refrain from moral wrongdoing against one's charges, even if one feels, or comes to feel, that one's children have violated this prohibition against oneself, for example, through their demands on one's attention, patience or resources that are felt increasingly as excessive and oppressive as they gain independence and maturity. If these obligations hold for the radical inequalities of power between infants and caretakers as well as for the roughly equal power among professional colleagues, they hold for all cases of moderately unequal power relations in between.

Now I think Baier overlooks the role of such obligations, not because of her Humean distaste for them, but rather because her analysis of trust focuses primarily on the trustor rather than on the trusted. Her primary concerns are what trust is; whom we can trust, and under what conditions; when our trust misfires, when it is unhealthy, irrational, or morally corrupt; and how to protect our ability to go on trusting in the face of repeated betrayals. That is, she works almost entirely from a perspective of epistemic ignorance as to when and whether trust is justified.

To answer these latter questions without appealing to justificatory principles is impossible. But when she turns her attention to principles of trustworthiness such as those Thomas Scanlon offers,[18] she rejects them as "unhelpful rules," "algorithms" (MP 160), and reliance on "our Kantian rational capacity to be law-abiders" (MP 151) as mere obstructions to the free-form, case-by-case use of our "powers of judgment" (MP 160). This builds on Baier's earlier attempt in *A Progress of Sentiments* to valorize judgment for

[18] Thomas Scanlon, "Promises and Practices," *Philosophy and Public Affairs 19* (Summer 1990), 199-226.

Hume as the "great enlarger of our reason" (PS 283; see generally PS 281-284). "Our capacity for judgment," she claims, "outruns our capacity to reduce our judgments to rule" (PS 281). But in fact judgment just is the application of general rules and principles to particular cases, and so part of what Baier needs to answer these questions. And her major objection to Scanlon's analysis is that our behavior provides no evidence that we accept his suggested principles (MP 134, 141-142), nor that we necessarily want anything so restrictive as can be encapsulated in any principle (MP 168-176).

But the pressing question is not what we in fact do or do not want or accept, any more than it is whether or not we *want* to do what we must to protect our ability to trust, or to contain the damaging effects of a Humean shame morality. My suggestion is that these two problems are related: We increase our ability to trust to the extent that we decrease the shaming consequences of openly acknowledging our moral vulnerability; and we do that to the extent that we meet our moral obligations to treat one another with tolerance, compassion, and respect, i.e. as ends in ourselves. The question, then, is rather what is *morally required* in order that a climate of trust can flourish, in which a rational and morally decent trustor can function.

Essentially, Baier's analysis answers the tactical question of how best to protect our moral innocence in an environment we must always fear is morally corrupt. Under the circumstances, that is an important and realistic question to try to answer. But if we want to know how, strategically, to create a moral environment in which such self-protective, blind groping for security is unnecessary, we need to know what reciprocal moral obligations we must voluntarily shoulder in order to realize that state of affairs. That we are required to fulfill certain basic moral obligations of trustworthiness, tolerance, compassion, and moral dependability in order that we each may feel reciprocally comfortable in exposing our weaknesses, flaws, dependencies, and moral vulnerabilities to one another is a fact that Baier's objections do not refute. So although her analysis of trust once again succeeds in making a persuasive case for reorienting the focus of normative ethics accordingly, it does not do so at the expense of the Kantian Contract-Theoretic model she purports to reject.

7. An Assessment of Baier's "Stylistic Experiment"

My treatment of Baier's version of Hume has so far not mentioned the very unusual, pervasive and disturbing style of exposition of *Moral Prejudices* – that element of it that best explains its title; and, in addition, most fully justifies my description of it in Section 1 as taking a quintessentially Humean indexical approach. Because this element figures very prominently in the experience of reading this book, I examine this aspect of it in some depth in this concluding section, and describe its effect on the reader. I suggest that Baier takes the indexical approach to doing philosophy far beyond the limits

Hume himself would have allowed; and that this, perhaps more than anything else, demonstrates Hume's (unlike Baier's) final allegiance to the very model of rationality – the Kantian model – he claims to repudiate.

I said in Section 1 that Baier challenges the standard, impersonal style of philosophical exposition with what she describes as a "stylistic experiment." Thus she argues that "[t]he impersonal style has become nearly a sacred tradition in moral philosophy, and examples of departure from it ... are not altogether encouraging .... The selective anecdote ... certainly has its own dangers, including those of bias. Nevertheless it seems to me time to experiment a bit with styles of moral philosophy, especially for those of us who, like Hume, hope that moral philosophy can be accurate without being 'abstruse' and might even 'reconcile truth with novelty'" (MP 194-5). From this passage we can infer that Baier's "stylistic experiment" will sometimes substitute personal remarks for impersonal judgments, selective anecdotes for extended analysis, and will run the risk of bias that attends such philosophically novel methods of argument. The passage occurs toward the end of *Moral Prejudices*, but by the time one reaches it one feels that its promise has been, unfortunately, more than fulfilled. The effect of Baier's stylistic experiment is to undermine the credibility of her arguments.

I have already cited some examples. There was her simplistic reduction of Kant's moral philosophy to rule-following, to "consulting some book of rules" (MP 54-55). And there was her *ad hominem* (and incidentally false, according to some of Kant's biographers) caricature of Kant (among others) as a misogynist and puritan bachelor "who had minimal dealings with women" (MP 114). Then there was her rush to judge Kant's "quality of mind and heart" (MP 273) for supporting *Lex Talionis*, without even considering the possibility that Kant's sympathy for and outrage on behalf of the victims of murder, rape, or pederasty might have had something to do with it; and her lurid characterization of Kant's ethics as sado-masochistic and "reeking of blood and torture" (MP 276-277), which might have been excusable coming from Nietzsche, who was, when he originally wrote this, probably starting to exhibit signs of the syphilitic insanity that would eventually kill him. And there was her immediate and unfounded speculation that Kant's "shocking" refusal to condemn unmarried mothers guilty of infanticide (MP 277) must have been motivated by cruelty.

But there are so many more instances of Baier's insulting jibes at Kant, and her perfunctory misrepresentations of his views, that it would be a waste of time to catalogue all of them.[19] She frequently conflates Kant's own views,

[19] I have done so nevertheless, to gratify the perversely titillated reader on the prowl, out slumming for lowlife philosophical exegeses: At MP 26 Baier dismisses Kant's concepts of rights, freedom and autonomy on the grounds that women's oppressors thought highly of autonomy, too, as though the trouble with suspect moral values was who held

the views of some particular Kantian such as Rawls or Scanlon, and some fictional collection of straw men – "the Kantians, or "contractarians" – whose purported views are too implausible to take seriously. Take, for example, her criticism that a Kantian Contract-Theoretic morality focuses merely on freedom under conditions of interpersonal conflict, whereas Hume attempts to solve the "deeper problem" of intrapersonal conflict and instability over time (MP 61). Clearly she cannot be leveling this criticism at Kant, who produced an epistemology articulating conditions for the unity of intuition,

---

them and how they were manipulated rather than what they were (also see MP 263 for more specious arguments that "Kant's much admired version of autonomy ... turns out on closer inspection to be a monopoly of a few representative propertied males"). At MP 30 she verbally cartoons a "Kantian picture of a controlling reason dictating to possibly unruly passions," which, I have already argued, was not Kant's view, nor that of any Kantian I know of. At MP 84-87 she gratuitously attacks Kant's categorical imperative procedure as isolationist, his concept of autonomy as selfish, and his concept of the kingdom of ends as inherently patriarchal and lacking "procedures for shared decision-making." At MP 88 she faults Kant (1) for his assumption that the same reasoning capacities exercised on the same subject matter under the same conditions will produce agreement among the reasoners, on the grounds that actual rational people do disagree, (2) for neglecting to explain how to judge only matters of interpersonal concern – as though the kingdom of ends formulation of the categorical imperative did not address this very question, and (3) for failing to free his mind of religious prejudice, as though he had not argued repeatedly and bravely, against governmental pressure, that God's commands must obey reason rather than vice versa. At MP 115 she claims that only philosophers who do not "remember what it was like to be a dependent child or who [do not] know what it is like to be a parent or to have a dependent parent, an old or handicapped relative, friend, or neighbor will find it [plausible] to treat such relations as simply cases of co-membership in a kingdom of ends," thereby adding the presumption of authorial omniscience with respect to various philosophers' memories to her caricature of Kant's kingdom of ends. At MP 248 she lampoons Kant's concept of the social relations among nations as "a sort of Leibnizian harmony of moral monads." At MP 250 she faults Kant for the "elitist or at least selective individualism" of his thought. Aat MP 255 and 257 she makes much of Kant's sexism, as though every male philosopher in the history of philosophy with the possible exception of Mill were not also patently sexist. At MP 277 she faults Kant for seeing a link between formulations and applications of the categorical imperative on the one hand and the mores of his own society on the other, as though this redounded to the discredit of the categorical imperative. And at MP 290 she suggests that "it might be easier for men than for women to be helped to self-definition by a reading of the full corpus of Kant's works," thereby managing to insult and confuse simultaneously all the women who have found it easy to be so helped and all the men who have found it difficult. Of course none of this prevents Baier from enlisting Kant in her own cause when this is convenient: As we have already seen, she uses a universalization argument to minimize the damage that would be done in a society governed by shaming principles, appeals to "rules and recipes" that may help us design lasting schemes of cooperation, and to the presence of a categorical imperative as one of the benchmarks of a genuine moral theory.

understanding and reason within the self, a moral philosophy that attempts to reconcile empirical inclinations with the demands of reason within the self, and an aesthetics that explores the relation between feeling and judgment within the self. And she cannot be thinking here of Rawls, who devoted a good third of his *Theory of Justice* to such questions of intrapersonal conflict as the relation between self-interest and impartiality, envy and equality, self-respect and just distribution, rationality, prudence, and moral development. So at whom, exactly, are these criticisms supposed to be directed? On this Baier offers no guidance.

Or consider her critique of Rawls's Kantianism. Baier reasons that since Rawls, who is "one of the ablest and most influential defenders" of the American liberal tradition, appeals to Kant and "sees Kant ... as giving us a moral and social philosophy which best articulates the basic principles of this nation's scheme of cooperation," she may turn to an examination of Kant "to understand what relationship there is, and is believed to be, between individual autonomous persons and the life that they live together under one constitution and one set of laws, and between individual and collective responsibility" (MP 253). Fair enough. But she does not examine that part of Kant's thought that influenced Rawls, i.e. primarily the *Groundwork of the Metaphysics of Morals*, or how Rawls himself interprets Kant.

Instead, Baier chooses to look at texts and passages in Kant's writings – in *Perpetual Peace* and the *Metaphysics of Morals* – that are largely irrelevant to Rawls's inspiration. So when she then complains that Kant does not, in those places, have much to say on the relation between individual and collective responsibility (MP 253), or that Kant's conception of republicanism underwrote the economic oppression of women (MP 255-257), or that Kant grants neither a legal nor a moral right of rebellion against tyranny (MP 258), and that "[t]his may seem an odd version of republicanism, and of the moral ideas behind it, to be taken as the model for a representative democratic republic such as this one, which began in a revolution. ... Can this be a variant of the republican ideal which inspired the Founding Fathers of this nation?" (MP 259), we no longer have any idea of to whom she is addressing these criticisms, or why.

Baier claims to want to "provoke the Kantians into explaining just how his ethics can escape the charges [she] make[s]" (MP xii). But the problem is that most of her charges are based on mischaracterizations too elementary to constitute a worthwhile challenge to Kantians. In such cases one wants suggest, rather, that she simply go back to the text and attend to it with the same care she gives to Hume – particularly since when she does so she rarely makes these mistakes.[20] Baier often remarks that she will leave it to others to

[20] See, for example, her sensitive treatment of the contrast between Kant and Hume on love and the dangers of intimacy (MP 34-36, 38, 39-42, 45); or her disapproving but

be fair to Kant (MP xii, 269, 290); but that is *her* job. She does not have to like Kant. But if she is going to discuss his views, it is her intellectual obligation to read him carefully and represent his views accurately to the best of her ability. The disrespect and lack of fidelity with which she often represents Kant's views express her philosophical distaste for him much more strongly than any objections – which are correspondingly unwarranted and unpersuasive – she actually levels against him. One would have thought she would have wanted not merely to express that distaste, but to give it a sound philosophical justification.

Then there is her stereotyping of women. Given that so many other European American, middle-class women academics have made the same mistake, it would be unfair to single out Baier for her embrace of Carol Gilligan's conclusions about female moral development, based, as they are, on a research sample that is so extremely provincial with respect to race and class.[21] But it is not unfair to expect from a major philosopher more careful and qualified empirical generalizations than that "[in] women's moral outlook, ... [there is] the tendency of the care perspective to dominate over the justice perspective in their moral deliberations" (MP 52); or that "women's [moral] theory, expressive mainly of women's insights and concerns, would be an ethics of love," whereas "men theorists' preoccupation [has been] obligation" (MP 4).[22] And it is not unfair to expect more searching and nuanced speculation than that maybe "reflective women, when they become philosophers, want to do without moral theory, want no part in the construction of such theories" (MP 2).

These ill-considered remarks about what "reflective" women are presumed to want are particularly objectionable. Baier says that Gilligan's research tells us how "*intelligent and reflective twentieth-century women* see morality, and how different it is from that of ... the men who eagerly assent to the claims of currently orthodox contractarian-Kantian moralities" (MP 115-116; emphasis added). So what of all the women philosophers engaged in Kant scholarship or in reconstructing a Kantian morality? What of the women political philosophers who not merely "eagerly assent" to but indeed assert and actively defend the claims of "contractarian-Kantian moralities"? Are we perhaps just not intelligent and reflective enough? Or trapped in the wrong century? Or not real women?

---

nevertheless fair treatment of Kant's analysis of beneficence and debt-avoidance (MP 190-191).

[21] Carol Gilligan, *In a Different Voice: Psychological Theory and Women's Development* (Cambridge, Mass.: Harvard University Press, 1982).

[22] Particularly since her normative views are so much more nuanced and sophisticated than these crude categorizations would suggest. See, for example, her recent "Note on Justice, Care, and Immigration Policy," *Hypatia 10*, 2 (Spring 1995), 150-152.

Of course Baier is, as we have already seen, no kinder to male philosophers, those "members of an all-male club" of "misogynists" and "puritan bachelors" concerned only to "read [their] *Times* in peace and have no one step on [their] gouty toes" (MP 114). To that Baier feels compelled to add that the preoccupation with Prisoner's Dilemma problems is a "big boys' game," and "pretty silly" (MP 2); that deontological principles are inherently "authoritarian" (MP 216) and "patriarchal" (MP 222); and that contract is a "male fixation" (MP 114).

In the very first pages of *Moral Prejudices* Baier notes the dangers of generalizing on such matters (MP 1, 2), and then declares that since "exceptions confirm the rule," she will "proceed undaunted" nevertheless (MP 2). But Baier's is the kind of generalizing that, when encountered in historical texts written by male philosophers, lead us to counsel our undergraduates on the challenge of reading philosophical texts carefully, and of patiently culling the deep philosophical insights from the chaff of personal prejudice and social anachronism that flaws every historical work. Seeing such personal prejudices expressed in print by a contemporary feminist philosopher concerned to fight against stereotypes of women certainly presents a pedagogical challenge, if no other kind.

But it is when Baier gets to her unfounded speculations about the biological limitations on women's professional potential that she inflicts the most serious damage on the women she means to support. She wonders whether "enough women professionally [will] survive their high estrogen years;" whether "they [will] be able to squeeze out enough articles while they are menstruating, gestating, and lactating;" and whether we should not expect them to hit their intellectual prime around age 50 when all that is behind them (MP 298) – forgetting, it seems, that that is roughly the age at which both Kant and Rawls began to produce their major philosophical contributions as well. Baier is, of course, entitled to her views about women and their intellectual and professional potential. But expressing in print her view of women as virtual slaves to their biology, whose capacity we must question, during their childbearing years, to do anything more than extrude various organic effluvia, including maybe an article or two every now and then, reinforces some of the ugliest and most damaging stereotypes – of women as cows, to put it bluntly – we have suffered. No male philosopher could get away with such pronouncements. Richard Brandt's notorious comparison of a hypothetical woman to a dog who is confused about what kind of feed it wants pales by comparison.[23]

It is impossible to doubt the depth and integrity of Baier's feminist commitment. So it is difficult to know what to make of all this. We have

[23] Richard Brandt, "Rational Desire," APA Western Division Presidential Address, *Proceedings and Addresses of the American Philosophical Association XLIII* (1969-1970), 43-64.

already seen that Baier explicitly rejects at least some of the traditional standards of philosophical analysis, and that she means to experiment with replacing extended impersonal analysis with a more personal, anecdotal style in a Humean spirit. When she flatly declines to say what she means by the word "cruel" at the beginning of her inquiry into whether Kant's ethics is more cruel than Hume's, she thereby rejects absolute adherence to the injunction that we try to instill in introductory philosophy courses to define one's key terms. Similarly, when she simply leaves in the text inconsistent statements, such as suggesting that trust in sustained trust may be a "supreme virtue" at MP 185 and denying that she intends to suggest this at MP 188, or suggesting at MP 168 that foundational trust can be spelled out in a moral principle, and at MP 182 that it cannot be, it is the principle of consistency itself – or perhaps merely Kant's injunction to *synthesize that manifold under rule-governed concepts!* – that she challenges. Here and elsewhere her stylistic experiment calls into question many of the most traditional and familiar standards that philosophers have taken for granted. In this respect she transgresses the rationalistic limits of Socratic metaethics that Hume himself so scrupulously observes.

I close this chapter by describing what I see as the dangerous consequences of Baier's stylistic experiment, by connecting it to her substantive project of putting the analysis and reparation of power imbalances at the center of normative ethics.[24] One characteristic of the discipline of philosophy in which we all can take pride is that there are identifiable professional standards of competence to which we are trained to adhere – standards we have inherited from Socrates himself. I spelled out these standards in Chapter I, but it does no harm to review them here. We are all, regardless of professional power or status, trained to discern when an argument is good or bad, consistent or inconsistent, superficial or searching, original or derivative, rigorous or sloppy, accurate or misleading, all regardless of the power or status of the individual who makes it. Most philosophers have a personal commitment to these standards independent of

[24] An earlier version of this chapter expressed my animadversions toward Baier's stylistic experiment much more vehemently and subjectively, in an attempt to convey how personally offensive and threatening I found her rejection of traditional philosophical standards. In part this was inspired by some sympathy with her challenge to the impersonal style – specifically, when it is exploited passive-aggressively to mask such personal animadversions under the guise of objectivity. This would be a legitimate criticism; and it may well be that to this brand of intellectual prevarication voicing one's biases openly and candidly, as Baier does, is the only antidote. Nevertheless, if one believes one's personal animadversions to be objectively warranted, then one ultimately undermines their objective force by framing them merely as personal prejudices. Whether or not Baier believes her philosophical views to be anything more than prejudices is at issue, as is what she effects by so often asserting them as matters of fact.

their own professional power or status in the field. Their ability to discern, independently of the professional repercussions of doing so, whether and to what extent an argument meets these standards, whether it be their own or someone else's, and regardless of the power or status of the individual who makes it, can be an important source of professional pride and self-confidence that more than outweighs the disadvantages of whatever power inequalities they may experience. Indeed, as I argued in Chapter I, most philosophers recognize the personal commitment to and application of these standards of competence as the great equalizer that makes professional power imbalances irrelevant to unbiased judgments of philosophical worth.

Of course not all individual philosophers invariably rely on these standards in making such judgments. Sometimes some individuals are too pressed for time to read a person's work carefully, or too far removed from the person's area of specialization to be confident of their ability to judge its worth accurately. In such cases, some may rely on other criteria they believe bear a lawlike relationship to philosophical worth, such as the person's educational pedigree, department ranking, class or ethnic background, gender, or race. And sometimes their high regard for these other criteria, which bear no lawlike relationship to philosophical worth, obscures or outcompetes the standards of philosophical competence that define it. The result is class-, gender-, or race-biased professional judgments that trade philosophical worth for sociocultural status; and thereby both drag down those standards of competence, and also reinforce the power imbalances that those standards of competence were supposed to equalize.

When an eminent philosopher of Baier's stature flouts these standards, her audience gets two messages: first, that she does not consider herself to be bound by them; and second, that she does not regard them as important. If the first message accurately represents her thinking, so much the worse for the quality of her work, regardless of the disciples she may attract. But if the second does, so much the worse for the quality of the discipline, whether one subscribes to the enterprise of Socratic metaethics or not. For the effect of the power, visibility and influence of her example is to ridicule, not only those standards, and the central intellectual values that make philosophy worth doing despite the professional corruptions of the field; but in addition those who view these standards and values as an important source of philosophical integrity and intellectual independence.

So it is not enough simply to observe that Baier often commits the genetic fallacy, or the inductive fallacy, or depends on *ad hominem* argument, any more than it would be enough merely to observe that she substitutes biased personal remarks for reasoned argument. These neutral observations do not address the public effect of her "experimental" methods. Baier rightly argues that we must trust others to discern our personal bias because we are so bad at discerning our own (MP 194). But we are not so bad at it that we cannot be

held to account for monitoring and managing our own personal biases at all. Baier does not need our help in discerning her personal biases, because as a philosopher and intellectual, trained to avoid such biases, she already knows what they are. By stating them explicitly nevertheless, without qualification or self-criticism, she effectively signals that it is not important to avoid them, and so that anyone who criticizes her on these grounds need not be taken seriously.

The consequence is that Baier thereby contributes, however unintentionally, to the progressive deterioration and devaluation of these standards; and so to a practice of professional evaluation that is less and less independent of those very power inequities – pedigree, connections, race and gender – she in her own trenchant analysis so eloquently deplores. She makes it permissible to dismiss someone's views based on name-calling or spiteful, negligent and misleading caricatures; and to applaud someone's views merely because they cleave to the preferred political ideology or are of the preferred gender or race. Her example licenses serious attention to such considerations of an argument's worth as, for example, whether the person who made it is a bachelor or a husband, childless or a parent, a man or a woman, of child-bearing age or past it – as though these had anything to do with its soundness. In effect she endorses such irrelevancies as legitimate refutations of an argument or view, and thereby, by implication, similarly irrelevant considerations for its acceptance, such as whether or not the person who made it is an upper middle-class, Anglo-Saxon, Protestant, highly pedigreed male in a top ten department. Her decision to flout traditional philosophical standards has the effect of undermining the very moral and political agenda she wishes to advance.

It is possible to interpret Baier's stylistic experiment as an attempt to fight fire with fire, i.e. to assault this entrenched but largely unspoken tradition by administering to it a dose of its own medicine. It would not be difficult to sympathize with the line of reasoning that the only way to put a stop to a longstanding tradition of gender-, race- and class-biased professional judgment is to turn the tables on its perpetrators, and give them an explicit taste of what it feels like to be on the receiving end, by mimicking it publicly. But this, too, would be ultimately a self-defeating strategy. For by participating in this degrading practice, Baier effectively legitimates it at the same time that she cheapens her own attempt to fight it.

What Baier's "stylistic experiment" proves most clearly, through its transgression of traditional philosophical standards of reasoning and analysis, is the absolute centrality to any recognizably philosophical discourse of the very criteria of transpersonal rationality it was Hume's primary objective to diminish. By carrying the indexical method so far beyond those of other Anti-Rationalists – indeed, to the point at which the method disintegrates into incoherent conversational sniping, Baier in effect closes the case in favor of its

Rationalist antithesis. So in the end, Baier not only does not make the case persuasively for Hume over Kant. She accomplishes exactly the opposite, by effectively conceding to Kant – or to a Kantian conception of the self – the very intellectual territory that set Hume and Kant at odds.

Chapter XIV. Hume's Metaethics

We now arrive at Hume's own pronouncements about the seemingly insoluble problems about moral motivation and rational justification with which this volume began. Up to this point I have said very little of substance about the transpersonal conception of reason that Humeans disparage as motivationally impotent and that a true Kantian would valorize as both motivationally and theoretically potent. I say a great deal about that conception in Volume II. But in discussing Hume's actual analysis it is necessary to preview that longer discussion, if only in broad outline.

The transpersonal conception of reason enfolds what I shall describe as a *traditional view*, according to which reason functions to, among other things, make inferences and categorical and hypothetical judgments, formulate hypotheses, and derive conclusions from evidential statements, deductive premises, and syllogisms. Reason on this traditional view is a logical arbiter, a calculator and discoverer of the relations between abstract concepts and states or events in the world. This is the very weak, conventional and widely accepted conception of reason to which Gewirth referred; and it is, as I have just argued in criticizing Annette Baier's interpretation of Hume, an important part of the transpersonal conception of reason on which the enterprise of Socratic metaethics, and indeed the practice of Anglo-American analytic philosophy more generally, relies.

Many have taken the utility-maximization model of rationality dissected in Chapters III and IV to be a direct consequence of the traditional view of reason. As we saw there, the utility-maximization model accepts the traditional view of reason as a purely theoretical or logical capacity, and assigns it the instrumental function of ascertaining, through investigation and calculation, the most efficient means possible of achieving our desired final ends, whatever these may be. Call this the *positive utility-maximization thesis*. Reason on the traditional view has two tasks, according to this positive thesis. Its primary task is to maximize utility; to discover the relations among phenomena such that they can best be utilized to satisfy our desires. Its secondary task is the examination of these phenomena themselves, for the purpose of discovering those objects or states of affairs that themselves best satisfy our desires. Such examination may run the gamut from methodologically rigorous scientific inquiry in general, i.e. the discovery of what phenomena there are, to a more restricted and informal scrutiny of particular objects, in order to discern or infer whether, or to what extent they have the qualities we desire. I call this task "secondary" because on the utility-maximization model of rationality, it is a special case of the primary task of reason, i.e. the utilization of our intellectual capacities in the service of realizing our desired final ends. Clearly the discovery of possible objects that

satisfy our desires is itself an end that reason may be used to achieve. Thus on this view we are thinking rationally if we successfully and appropriately perform those intellectual operations characteristic of theoretical reason on the traditional view. We are acting rationally if we successfully deploy these operations in realizing our desired final ends.

We saw in Chapter III that one immediate implication of the utility-maximization model of rationality as I have stated it is that reason has nothing to say about whether one's desired final ends themselves are rational; thus were generated the problems of rational final ends and moral justification discussed earlier in this volume. Call this the *negative utility-maximization thesis*. Like the positive utility-maximization thesis, this negative one does not follow from the traditional view of reason, but it does follow from the utility-maximization model of rationality. For this model regards reason itself as nothing more than a means for achieving our ends. Of course reason, on this view, may enable us to discover what ends we genuinely want, and may enlarge the scope of ends from which we may choose. But as we have seen in discussing Frankfurt, Watson, Williams, Slote, Rawls, Brandt, and others, it provides no criteria for identifying those ends themselves as rational, independently of their efficiency as means for promoting further ends to which they may be subordinate. Reason functions solely as the unique second-order means for determining the logical or material first-order means to our ends, whatever they may be. Together the positive and negative theses constitute an informal characterization of the utility-maximization model of rationality. I argue here that Hume himself embraces both theses; therefore the utility-maximization model of rationality in its entirety; and therefore the belief-desire model of motivation that furnishes its conative force.

Although this model is first articulated in Hobbes' *Leviathan*, the positive and negative theses of the utility-maximization model of rationality receive their first detailed explication and justification in Hume's *Treatise*[1], and the negative thesis is defended most forcefully there. Hume's most celebrated passages include those in which he characterizes reason as nothing but the "slave of the passions" (T 415), and as wholly silent on the question of whether I should "chuse my total ruin, to prevent the least uneasiness of an *Indian* or person wholly unknown to me. 'Tis as little contrary to reason," Hume continues, "to prefer even my own acknowledg'd lesser good to my greater, and have a more ardent affection for the former than the latter" (T 416). These

[1]David Hume, *A Treatise of Human Nature*, Ed. L. A. Selby-Bigge (Oxford: Clarendon Press, 1968). Henceforth all page references to the *Treatise* will be parenthecized in the text, preceded by T. All paragraph/page references to the second *Enquiry* (David Hume, *Enquiry Concerning the Principles of Morals*, Ed. L. A. Selby-Bigge (Oxford: Clarendon Press, 1966), will also be parenthecized in the text, preceded by E.

claims certainly seem counterintuitive in the ways earlier described, and commentators on Hume have not been happy about taking them at face value. We are often told that Hume took a perverse pleasure in attention-getting hyperbole,[2] and that we should therefore take them with a grain of salt. However, if the only evidence given for Hume's putative perversity were the passages we are instructed to disregard, it would not be evidence enough; nor would it be consistent with the honorable convention of showing respect for a thinker by assuming that she means what she says. But Annette Baier is not alone among those of Hume's commentators who have attempted the more ambitious project of finding positive and substantive evidence that Hume did not mean what he said in these passages; of fashioning a more constructive account of reason's role in constraining us to rational final ends elsewhere in the *Treatise*; and hence of showing that the many objections to the negative utility-maximization thesis discussed in Chapters VIII and IX are misplaced.

However, I do not agree with these more charitable interpretations of Hume. I argue here that a detailed reconstruction of Hume's arguments on these matters does not support these well-intentioned defenses of Hume; that he means *exactly* what he says in the controversial passages, and therefore embraces the utility-maximization model of rationality wholeheartedly; and consequently, that the many objections discussed in the foregoing chapters of this volume must be allowed to stand. I begin in Section 1 by demonstrating that on the face of it at least, Hume's view of rationality is straightforwardly identifiable as the utility-maximization model. I then argue in Sections 2 and 3 that this is fully consistent with his larger project of denying the motivational efficacy of reason. Sections 4 and 5 are devoted to elaborating in considerable detail a particularly compelling version of an argument claiming to show that Hume does impose restrictions on the range of final ends identifiable as rational, and Section 6 to refuting that argument.

1. Hume's Model of Reason

That Hume accepts the traditional view of reason described above is not difficult to ascertain. His conception is first introduced in Book I of the *Treatise of Human Nature*, where he divides reason into three kinds: (1) knowledge,

[2]See, for example, in addition to Baier's discussion in *A Progress of Sentiments*, Henry David Aiken, "An Interpretation of Hume's Theory of the Place of Reason in Ethics and Politics," *Ethics 90* (October 1979), 68; D. D. Raphael, "Hume's Critique of Ethical Rationalism," in William B. Todd, Ed. *Hume and the Enlightenment* (Edinburgh: The University of Edinburgh Press, 1974), 19; David Fate Norton, *David Hume: Common-Sense Moralist, Sceptical Metaphysician* (Princeton: Princeton University Press, 1982), 100; David Miller, *Philosophy and Ideology in Hume's Political Thought* (Oxford: Clarendon Press, 1981), 40, 47.

which he describes as a feeling of certainty or assurance produced by the comparison of ideas, (2) proofs, or arguments derived from causal relationships about whose soundness we feel no doubt or uncertainty, and (3) probability, which is that evidence about which we continue to feel uncertainty. Probability is then subdivided into chance, which Hume defines as the negation of a cause, and causes, which he characterizes as a constant conjunction of events which produces in us a habit of associating the idea of the one with the idea of the other (T 124).[3]

However, categories (2) and (3) partly collapse into each other, for Hume has earlier argued that certainty arises solely from the comparison of ideas and the discovery of unalterable relationships such as resemblance, proportion in number and quantity, contrariety, etc.; and that none of these are implied in the claim that whatever has a beginning has a cause (T 79). Causal relationships are therefore neither intuitively nor demonstrably certain. Therefore nothing, in point of fact, satisfies Hume's description of a proof (2); and causal relationships are a species of probability. This conclusion is partly confirmed by Hume's claim, a few pages later, that

> [t]he gradation ... from probabilities to proofs is in many cases insensible; and the difference betwixt these kinds of evidence is more easily perceived in the remote degrees, than in the near and contiguous (T 131).

Hence the basic categories of reason are knowledge, consisting in the comparison of ideas which gives rise to a feeling of certainty, and probability, i.e. that uncertain evidence arising from the observing of actual events.

Hume's later treatments of reason change his terminology but not this basic twofold division. In Book II, Section III ("Of the Influencing Motives of the Will"), Hume distinguishes between abstract or demonstrative and probabilistic reasoning (T 413-14). The first concerns only the abstract relations of ideas, which we may assimilate to Hume's earlier description of knowledge as consisting in the comparison of ideas – category (1) above; the second consists in an inquiry into the relationship between and among objects of experience, i.e. their causal relations – category (2)/(3) above. And as we have already seen, causal relations can be ascertained only with varying degrees of probability. This is then later confirmed, by implication, when Hume characterizes reason as consisting in two basic operations of the understanding: (1′) the comparing of ideas; and (2′) the inferring of matters of fact (T 463).[4]

---

[3]This characterization of what I call the "traditional view" is, I think, consistent with what Barbara Winters describes as the "naturalistic conception." See her "Hume on Reason," *Humes Studies V*, 1 (April 1979), 20-35.

[4]Thus I find no evidence for David Miller's contention that in Book II, Hume uses the term "reason" to cover all the operations of the understanding, including imagination,

Hume also characterizes reason as the discovery of truth or falsehood. This consists in the agreement or disagreement to the actual (Hume uses the term "real") relations of ideas, or to actual existence and matters of fact (T 458). Hume's intent in this passage is to argue that our actions, passions, and volitions can disagree with neither. What *can* agree or disagree, either with the real relations of ideas or with real existence and matters of fact, i.e. what can conform to reason in this way? Hume has already argued that this role is filled by our prior, unreflective ideas and impressions (T 415). These must conform or fail to conform to the ways in which ideas or events are in fact related.

Hume charts the relations between demonstrative and probabilistic reasoning in Book II, Section III of the *Treatise*, and there we find the relation to be essentially one of means to ends:

> Mathematics, indeed, are useful in all mechanical operations, and arithmetic in almost every art and profession: But 'tis not of themselves they have any influence. Mechanics are the art of regulating the motion of bodies to *some design'd end or purpose*; and the reason why we employ arithmetic in fixing the proportions of numbers, is only that we may discover the proportions of their influence and operations. ... Abstract or demonstrative reason, therefore, never influences any of our actions, but only as it directs our judgment concerning causes and effects, which leads us to the second operation of the understanding (T 413-14; italics in text).

Hume then goes on to describe how, when we are confronted by an object that causes us pleasure or pain, we feel an attraction or aversion to it, which in turn makes us "cast our view on every side, comprehend[ing] whatever

---

judgment, and belief (Miller, pages 40 and 47; *op. cit.* Note 2). Miller earlier refers to the passage in the *Treatise* in which Hume states that "[w]hen I oppose the imagination to the memory, I mean the faculty, by which we form our fainter ideas. When I oppose it to reason, I mean the same faculty, excluding only our demonstrative and probable reasonings" (T 117-18). Miller remarks on this passage that "[i]n the last sentence 'reason' is expanded to include the rule-governed imagination, which forms all 'probable' judgments (i.e. judgments concerning matters of fact not immediately present to the senses), and contrasted with the 'fanciful' imagination. In seeking to eliminate one source of confusion, Hume has inadvertently introduced another (the broader sense of 'reason' is frequently used by Hume in expounding his moral philosophy)" (Miller, page 27 fn.). But I fear the muddle here is not Hume's. Surely Hume means to say that imagination is the faculty by which we form our fainter ideas *except for our demonstrative and probable reasonings, which are formed by the faculty of reason.* Presumably the point of the contrast between reason and imagination is to distinguish between those faint ideas which are formed by nonrational mental processes and those which are formed by "demonstrable and probable reasonings." I do not see that Hume has expanded his use of the term "reason" at all.

objects are connected with [the] original one by the relation of cause and effect" (T 414).

Thus Hume's conception of reason is a hierarchically structured series of means to the ends we adopt. At the top of the hierarchy, we find abstract or demonstrative reasoning: the comparison of abstract ideas that characterizes mathematics and arithmetic. But abstract reasoning is merely a means enabling us to calculate probabilities more accurately. At the second level in the hierarchy, then, we find probabilistic reasoning: that brand of calculation that is concerned with causal relations between events. However, this too is merely a means to the further end of pursuing pleasurable objects and avoiding painful ones. We thus find this goal at the first and bottom level of the hierarchy, for it itself is not a means to any further end. The general appetite to good (or pleasure) and aversion to evil (or pain)[5] "arise originally in the soul, or in the body, whichever you please to call it, without any preceding thought or perception" (T 276). So for Hume, abstract reasoning is a means to probabilistic reasoning; probabilistic reasoning is a means to the rational manipulation of empirical conditions; and this in turn is the means to the objects of our desires. Hume not only accepts the traditional view of reason as essentially inference and calculation, but also, apparently, the positive utility-maximization thesis.

Hume makes his adherence to this thesis clear in a number of places. Directly after limning his hierarchical picture of reason, he goes on to explain how, when we incline or are averse to some particular object, based on the amount of pleasure or pain we expect from it, we utilize our reason in order to discover the causal relations that lead to or away from it, and design our actions accordingly:

> Here then reasoning takes place to discover this relation; and according as our reason varies, our actions receive a subsequent variation. ... It can never in the least concern us to know, that such objects are causes, and such other effects, if both the causes and effects be indifferent to us (T 414).

Later, in Book III, Section I ("Moral Distinctions Not Deriv'd From Reason"), Hume articulates this view of reason's function even more explicitly:

> [R]eason, in a strict and philosophical sense, can have an influence on our conduct *only* after two ways: Either when it excites a passion by informing us of the existence of something which is a proper object of it; or when it discovers the connexion of causes and effects, so as to afford us the means of exacting any passion (T 459; emphasis added).

[5]Hume identifies them at T 276 and 439.

That reason can function only as a means to achieve objects we desire, either by alerting us to the existence of such objects, or by charting the causal path to their attainment, implies not only Hume's acceptance of the positive utility-maximization thesis, but indeed the negative one as well. For that reason can only be a means to our ends clearly implies that it does not function to circumscribe those ends themselves.

Both of these theses are buttressed further by Hume's claims in the *Enquiry Concerning the Principles of Morals*. His adherence to the positive thesis is supported by his claims that

> nothing but [reason] can instruct us in the tendency of qualities and actions, and point out their beneficial consequences to society and their possessor (E 234/285).

and that it

> directs only the impulse received from appetite or inclination, by showing us the means of attaining happiness or avoiding misery (E 246/294).

As in the *Treatise*, Hume is quite explicit on the point that, just as reason discovers causal means for the realization of particular ends, similarly reason itself is the means by which we discover those causal relationships most suitable to their attainment.

Hume is most explicit in his affirmation of the negative utility-maximization thesis in the *Enquiry*. There he maintains quite clearly that

> the ultimate ends of human actions can never, in any case, be accounted for by *reason*, but recommend themselves entirely to the sentiments and affections of mankind, without any dependence on the intellectual faculties (E 244/293; italics in text).

Similarly, he argues that we require the sentiment of humanity, i.e., a feeling for the happiness of mankind and a resentment of their misery, in order to be motivated to promote these ends; for

> were the end totally indifferent to us, we should face the same indifference to the means ... *reason* instructs us in the several tendencies of actions, and *humanity* makes a distinction in favor of those which are useful and beneficial (E 235/286; italics in text).

In both passages the point is the same: It is not reason, but rather our passions and sentiments, which determine the ends that reason helps us achieve. Thus Hume's view satisfies the two essential conditions of the utility-maximization model of rationality.

2. Hume's Model of Motivation

The view I have attributed to Hume can be understood in two ways, and the discussion so far has emphasized only one of them. I have been concerned

to show that for Hume, there can be no conception of rational final ends, i.e. ends that conform to the prescriptions of reason. This is because Hume's utility-maximization model of rationality issues no such prescriptions. Its purview is confined solely to the discovery of means to those ends, and imposes no criteria of rationality on those ends themselves. The passages adduced so far seem clearly to point to this conclusion. But Hume's intention was more comprehensive. He wanted to show not only that reason could not determine rational ends of action, but also that it could not motivate action either. That is, Hume defended the belief-desire model of motivation.

This project was fueled by an interest in refuting the position, championed by Samuel Clarke and William Wollaston, that the conclusions of theoretical reason – i.e. the capacity to analyze and to perform logical operations – concerning the meaning of moral propositions were sufficient to incite one to morally virtuous action. Samuel Clarke offered an analysis of morally right actions as those which are self-evidently fitting or suitable to the circumstances in which they occur. This suitability or fitness is generated by natural proportional relations and uniformities that obtain among natural objects and events, just as they do among geometrical and mathematical entities. Hence, he argued, it is self-contradictory to will acts that are recognized to be unsuitable to their circumstances, i.e. immoral.[6] Wollaston, on the other hand, rejected Clarke's analysis of rightness as fittingness. Instead he held that moral actions are those which assert logically true propositions, while immoral actions are self-contradictory.[7] Thus his conception of moral rightness is equivalent to that of truth. However, both Clarke and Wollaston concurred in the belief that these convictions were discoverable *a priori* by theoretical reason, i.e. that a simple examination of the nature of action and its circumstances would reveal those actions which were morally right. And significantly, both believed that mere recognition of these "moral facts" placed the agent under obligation to act in conformity with them.[8]

Against this view, Samuel Clarke's two foremost critics, John Clarke and Francis Hutcheson, argued that moral propositions did not analyze the nature of moral action, but rather were concerned with moral obligation. For since the mere recognition of fittingness or self-consistency had no conative force,

---

[6]Samuel Clarke, *A Discourse Concerning the Unchangeable Obligations of Natural Religion*, Ed. L. A. Selby-Bigge, *The British Moralists, Vol. II* (New York: Dover, 1965), 4-6. See the discussion of Clarke by Rachel Kydd, *Reason and Conduct in Hume's Treatise* (New York: Russell and Russell, 1964), Chapter I.

[7]William Wollaston, *The Religion of Nature Delineated*, in Selby-Bigge, *ibid*., 362-4. See Kydd, *ibid*.

[8]Clarke, *op. cit*. 12-14, 16-17, 23-4, 31-3; Wollaston, *ibid*. 370-1; Kydd, *op. cit*. 28-36.

such propositions could not move an agent to do or refrain from any act, hence could not be central to a true analysis of moral propositions.[9] The central topic of moral philosophy was thus what we are obligated to do. Conflating what we are obligated to do with what we are compelled or *obliged* to do, both then conclude that an action cannot be called obligatory unless the agent feels impelled to perform it. So either reason had to be rejected as the source of morality, or else reason itself had to discover its own special motive to action.[10]

Hutcheson is clearest on this latter requirement, and most pessimistic about its fulfillment. He maintains that we can only be moved to action by "exciting reasons," and these are dependent on our desires. But (as Kydd points out) since our desires are empirical, *a priori* rational analysis cannot of itself incite us to action:

> As if indeed reason, or the knowledge of the relation of things, could excite to action when we proposed no end, or as if ends could be intended without desire or affection.[11]

Hutcheson's claims bear further consideration. His point in this passage is twofold. First, rational *a priori* analysis bears no relation to desires and emotions, and only these can motivate us to action. But second, the reason theoretical reason fails to move us to action is not only because it is neither a desire nor an emotion. It fails because it provides us with no end about which we might be able to feel a desire or aversion or emotion. So even if theoretical reason could fashion some object proved by analysis to be ultimately worthwhile (such as Kant's highest-order, transcendent ideas of God, freedom and immortality), this would be irrelevant to the moral enterprise if it were not the object of a desire. Hence desires and affections are not significant merely because they move us to act; impulses, whims, and uncontrollable urges do so as well. Desires are significant because they posit ends that we desire to achieve, and that therefore move us to try to achieve them.

Two implications of Hutcheson's argument follow directly. First, a necessary condition of an object's having moral value is that it be able to motivate us to action, i.e. that it be an object of desire. Second, reason provides no such motivating ends. The conclusion is clear: Reason provides no moral motivation to action. But if reason provides no motivating ends, and if we can be motivated only by ends we desire to achieve, then reason does not determine the ends we desire to achieve; these can be determined only by

---

[9]Kydd, *op. cit.* 23.
[10]Kydd, *op. cit.* 38.
[11]Francis Hutcheson, *Illustrations on the Moral Sense*, Ed. Bernard Peach (Cambridge, Mass.: Belknap Press of Harvard University, 1971), 122.

instincts, affection, and desire.[12] This conclusion is recognizable as Hume's negative utility-maximization thesis.

As with Hume, this negative thesis is buttressed by Hutcheson's answer to the question,

> [A]re there no exciting reasons, even previous to any ends, moving us to propose one end rather than another? To this Aristotle long ago answered that 'there are ultimate ends desired without a view to anything else.' To subordinate ends those reasons or truths excite, which show them to be conducive to the ultimate end, and show one object to be more effectual than another; thus subordinate ends may be called reasonable. But as to these ultimate ends, to suppose exciting reasons for them, would infer that there is no ultimate end, but that we desire one thing for another in an infinite series.[13]

Here Hutcheson does not mean to deny that we are motivated to achieve final ends. Rather, he is denying that we are motivated by rational considerations to achieve those ends. His point is that reason plays no role in the choice of final ends. Furthermore, reason does play a role in investigating and determining the most effectual subordinate ends, i.e. means to those final ends.[14] This view is recognizable as Hume's positive utility-maximization thesis.

Thus Hume's task was twofold. First, it was necessary to clearly delineate the actual scope and limits of reason, in order to demonstrate conclusively the conviction he shared with Hutcheson and John Clarke that no truth of reason could of itself incite an agent to action, much less moral action. Second, Hume had to provide a positive and detailed account of the passions in order to show just what the true origins and motives of moral action actually were. These enterprises form most of the subject matter of Books II and III of the *Treatise of Human Nature*, and account for his adherence to both the positive and the negative utility-maximization thesis.

For it is of course significant that both Hume and his ally Hutcheson assume almost without a second thought the truth of the negative utility-maximization thesis as an argument supporting their convictions about

---

[12]This point is supported, and not undermined, as Kydd seems to think (*op. cit*., 39-40), by his later assertion that

> He acts reasonably, who considers the various actions in his power, and forms true opinions of their tendencies; and then chooses to do that which will obtain the highest degree of that to which the instincts of his nature incline him (*ibid*. 126).

[13]*ibid*., 123.

[14]Cf. Note 17 and also *ibid*., 115-16, where he describes reason as the "sagacity in prosecuting any end," and as the finding of means to promote both the public and private good.

reason's irrelevance to moral, and in general behavioral, motivation. Both suppose that reason's inability to determine rational ends, and its limited function as a mere means for achieving those ends are in some sense indicative of its inability to motivate an agent to action. The implicit reasoning seem to be that a necessary condition of motivation is an object of desire, which they equate with an end, and that if reason cannot determine such an end, it cannot move one to action. Hutcheson follows this line of reasoning straightforwardly: He argues that a rational end is a necessary condition of rational motivation, and that since our ends are ultimately determined by our nonrational desires, this condition cannot be satisfied.

Hume's strategy is subtler, and more problematic. His objective is to demonstrate the mutual independence of reason and motivation. But as we shall see in Section 3, his arguments depend on confusing a motive and an end of action. This confusion then leads him to conclude, from the imperviousness to rational standards of certain ends, to the imperviousness to reason of our motives for acting – just as Hutcheson does. However, this thesis will need to be evaluated independently of Hume's arguments, for they do not prove what he thought they did.

Hume begins by considering the role of passion, and then later makes the role of reason his starting point. The rest of this section will be devoted to the first, and Section 3 to the second. His first argument, then, is that reason cannot incite us to action. Only the prospect of pleasure or the avoidance of pain from an object can do that (T 414). He states quite clearly that reason has no motivational efficacy (T 415), and later characterizes it as "of itself ... utterly impotent [to excite passions, and produce or prevent actions]" (T457). Moreover, in his own summation of his argument of Book II, Part III, Section 3, Hume takes himself to have "prov'd, that reason is perfectly inert, and can never either prevent or produce any action or affection" (T 458). Nor can reason oppose our desires, for only another desire or passion can oppose a desire or passion, and if this could originate in reason, then reason would, on the contrary, be capable of inciting us to action. And Hume has just argued that it is not. Thus Hume's first claim is that only passions can oppose each other, and only passions can motivate actions. Reason, it seems, is excluded from the scene.

However, there are passages in the *Treatise* that have seemed to many to commit Hume to at least some minimal motivational role for reason, and these must be examined. First, there are the passages in Part III, Section 10 of Book I, "Of the Influence of Belief" on which we saw that Baier relies. There Hume tells us, for example, that "the effect ... of belief is to raise up a simple idea to an equality with our impressions, *and bestow on it a like influence on the passions*" (T 119; emphasis added). He also states that "belief is almost

absolutely requisite to the exciting our passions" (T 120). The implication would seem to be that belief constitutes an identifiable link in the causal chain between the presence of the object and the agent's exertion in its service. If belief is motivationally influential in exciting the passions, which in turn cause action, then to the extent that true belief satisfies criteria of reason, reason must be capable of motivational influence as well.

However, one of the premises contained in this line of reasoning is subject to doubt: belief may be motivationally influential, but not even true belief is a species of reason for Hume. To see this, consider first Hume's detailed account of how facts become "the object of faith or opinion":

> When any affecting object is presented, it gives the alarm, and excites immediately a degree of its proper passion; ... This emotion passes by an easy transition to the imagination; and diffusing itself over our idea of the affecting object, makes us form that idea with greater force and vivacity, and consequently assent to it, according to the precedent system (T 120).

The steps in the process are (1) the affecting object causes a passion; (2) this passion is transferred to the imagination; (3) in the imagination, the passion infuses our idea of the object; (4) this infusion imparts greater force and vivacity to the idea, "imitating," as Hume has said shortly before, "the effects of the impressions;" (T 119); (5) the greater intensity of this idea, and its approximation to an impression causes us to assent to it. "Belief," Hume tells us, "is nothing but *a more vivid and intense conception of any idea*" (T 119-20). The implications are four. First, belief is composed of an idea and a passion "diffused over" it. Second, the causal factor in belief is the passion *that precedes the idea it infuses*, not the idea itself. Third, since reason, as we already know, concerns only relations of ideas and matters of fact, reason is no more causally efficacious than are ideas as such. And finally, therefore, belief, *qua* passion-infused idea, is not a species of reason.

This account of the influence of belief is borne out by Hume's earlier analysis of the nature of propositional belief in Sections 6 and 7. There Hume distinguishes between belief in those propositions proved by intuition or demonstration, and those concerning causation and matters of fact (T 95). We are determined to believe the former either immediately or by the interposition of other ideas. This chain of ideas, i.e. inference, depends solely on the union and association of ideas in imagination, not on reason (T 92). By contrast, whether we believe a proposition about matters of fact or its negation is determined by which of the two ideas is related to or associated with a present impression, thus increasing its force and vivacity (T 96; also T 86, 93). As Hume frequently reminds us, belief is a particular *manner* of forming an idea (T 95, 96, 97). A belief that has motivational influence, then, is an idea whose accompanying impression has sparked the passion that infuses

it and has thereby rendered it particularly forceful and vivacious. Again it is the impression and the passion preceding the idea that are motivationally efficacious, not reason.

This conclusion is further supported by Hume's claims that belief is merely a certain feeling or sentiment (T 153, 624); that it is not itself an idea (T 184, 623-26) or a simple act of thought (T 184); and that it is more properly an act of the sensitive than the cognitive faculties (T103, 183-5). Hume in the *Enquiry* makes the point even more strongly: He characterizes belief as "the true and proper name of [an indefinable sentiment or] feeling" (E 40/48-9); he contends that

> [B]elief consists not in the peculiar nature or order of ideas, but in the *manner* of their conception, and in their *feeling* to the mind. I confess, that it is impossible perfectly to explain this feeling. ... But ... we can go no farther than assert, that *belief* is something felt by the mind, which distinguishes the ideas of the judgment from the fictions of the imagination. It gives them more weight and influence; makes them appear of greater importance; enforces them in the mind; and renders them the governing principle of our actions (E 40/49-50; italics in text).

These passages lend support to the thesis that what identifies something as a belief is the passion that imbues it, not the idea that gives it content. Having come to believe something, it may well be that our believing it causally influences the passions that cause us to act. But the source of its causal influence is the passion that infuses it; this in turn influences the passions that directly cause us to act. It is thus false, according to Hume's account, to infer that reason itself has any such influence on action.

However, there are two other sets of passages that may seem to engender similar inferences. Hume often claims that reason *alone* cannot influence the will (T 413, 414, 457); that reason can "excite" a passion only "by informing us of the existence of something which is a proper object of it" (T 459); that an action "may be *obliquely* caus'd by [a judgment], when the judgment concurs with a passion" (T 459; italics in text); that reason "may, indeed, be the mediate cause of an action, by prompting, or by directing a passion" (T 462); and that "the blind motions of the [affections], without the direction of the [understanding], incapacitate men for society" (T 493). These passages have suggested to some that reason may be at least a necessary (if not sufficient) motivational influence on a passion.[15]

---

[15]See Henry David Aiken, "An Interpretation of Hume's Theory of the Place of Reason in Ethics and Politics;" and David Fate Norton, *David Hume: Common-Sense Moralist, Sceptical Metaphysician, op. cit.* Note 2. As far as I can tell, W. D. Falk (in "Hume on Practical Reason," *Philosophical Studies 27* (1975), 1-18) does *not* make this mistake.

Here the problem lies in the scope of the word "cause" as we, and Hume, choose to use it. It would seem that Hume, and some of his commentators, have failed to make the distinction between a necessary condition and a contributing cause. Something is a *necessary condition* for an action if the action would not have been performed without it. Something is a *contributing cause* of an action if, independently of other causal factors with which it is conjoined, it exerts some causal influence on the agent to perform the action. There is no necessary connection between necessary conditions and contributing causes of some event. Suppose, for example, that I discover a craved cherry pie on the table. I am moved to approach the table. Does my discovery of the pie move me toward the table? Surely not. If I discovered the pie without wanting it, it would have no such influence. Rather, it is my desire for the pie that has this effect on me. Of course my discovery of the pie on the table is a necessary condition of my approaching the table (rather than, say, the window). In that sense, my discovery "directs" or "prompts" me toward the table. But not everything that is required in order for an event to occur can be sensibly described as a contributing cause of its occurrence.[16] In particular, my discovery of the pie is a necessary condition of my action, but not a contributing cause of it; for, as Hume often notes, reason by itself has no causal influence whatsoever.

The suggestion, then, is that when Hume uses words such as "prompts" or "directs", he is referring to a particularly salient necessary condition of action, i.e. reason – not a contributing cause of it. This interpretation enables us to resolve the passages just cited with Hume's immediately preceding claim to have "prov'd, that reason is *perfectly inert*, and can never either prevent or produce any action or affection" (T 458; emphasis added). That Hume regards these two points as mutually consistent is made clear in the *Enquiry*, when he states that

> [r]eason being cool and disengaged, is no motive to action, and directs only the impulse received from appetite or inclination, by showing us the means of attaining happiness or avoiding misery (E 246/294).

That reason directs or prompts, then, does not imply that reason motivates; quite the contrary, on Hume's account.

Finally, there are the passages surrounding Hume's account of the origin of the artificial virtue of justice. Hume tells us that society is advantageous for the purpose of compensating individual defects, achieving equality or

[16] Of course some contributing causes of action are necessary conditions, as Hume recognizes (E 76/60). But in this passage he accords no higher priority to this quasi-nomological definition of cause than he does to his own inductive one offered later in the same paragraph.

superiority relative to others, augmenting individual abilities, and providing personal security (T 485) and protection of personal goods (T 488). But, he adds, "in order to form society, 'tis requisite not only that it be advantageous, but also that men be sensible of these advantages" (T 486), and that they gain this sensitivity from experiencing a family. On the other hand, our innate selfishness and partiality works against the cooperation with others that enables society to perform this role. "From all which it follows," Hume concludes, "that our natural uncultivated ideas of morality, instead of providing a remedy for the partiality of our affections, do rather conform themselves to that partiality, and give it an additional force and influence" (T 489). Where might we find a remedy for the partiality of our affections? Hume's answer follows:

> The remedy, then, is not deriv'd from nature, but from *artifice*; or more properly speaking, nature provides a remedy in the judgment and understanding, for what is irregular and incommodious in the affections (T 489).

Some commentators[17] have taken Hume to mean here that reason compensates for the partiality of the affections, hence provides a more stable source of motivation than they alone could supply. But first, this is not what Hume means; and second, even if it were, it would not imply that reason had motivational influence. That Hume does not mean to identify reason as the remedy for our partiality is suggested by his characterization of the remedy as "deriv'd from *artifice*"; reason, surely, is not derived from artifice. But Hume's real meaning can be seen more clearly by his subsequent remarks in the same paragraph: He explains that the remedy for social disturbance must consist in "putting [external goods], as far as possible, on the same footing with the fix'd and constant advantages of the mind and body," so as to limit "their looseness and easy transition from one person to another." "This can be done," he avers,

> after no other manner, than by a convention enter'd into by all the members of the society to bestow stability on the possession of those external goods, and leave every one in the peaceable enjoyment of what he may acquire by his fortune and industry (T 489).

The remedy for our partiality, then, is not reason, but rather the rules of justice, which ensures social equilibrium by enforcing the rules of private property. At most, reason is the source of the rules we devise for this purpose. Thus to say that nature provides a remedy *in* the judgment and understanding is not to say that nature provides the judgment and understanding *as* a remedy. Hume asserts the former, but not the latter.

[17]For example, David Fate Norton (page 134, *op. cit.* Note 2).

But suppose reason were Hume's remedy for the partiality of our affections? Would this show that it had motivational influence? I think not, for Hume makes it quite clear, in this paragraph and in the subsequent discussion, that we devise and implement the rules of justice for purely instrumental reasons, i.e. so that we may each enjoy our possessions in peace and security:

> By this means, every one knows what he may safely possess; and the passions are restrain'd in their partial and contradictory motions. *Nor is such a restraint contrary to these passions; for if so, it cou'd never be enter'd into, nor maintain'd; but it is only contrary to their heedless and impetuous movement* (T 489; emphasis added).

Clearly Hume means to deny any suspected departure from his earlier doctrine regarding the slavish and purely instrumental role of reason relative to the passions. Reason, under the guidance of self-interest (T 492), generates the rules of justice as means for restraining the passions, which in turn is the means to the safe enjoyment of property. Reason does not causally oppose the passions, but rather directs them in the sense noted above (T 493). Hence the role of reason in engendering the rules of justice is not only fully consistent with Hume's doctrine of Book II regarding reason's motivational inefficacy; it is an instance of that doctrine. We have yet to find the clear evidence of a conflicting doctrine upon which some of Hume's commentators have insisted.

3. The Passions from the Viewpoint of Reason

Recall that Hume's doctrine of the motivational inefficacy of reason was the first of two lines of thought, the first taking the viewpoint of the passions, the second taking the viewpoint of reason. Now let us consider this second line of attack more closely. Reason, as Hume has already established, consists in the conformity to truth, either of abstract relations between ideas or of experienced matters of facts, of our previous ideas and impressions. The passions, on the other hand, are neither. They are "original modifications of existence" that do not represent anything, and therefore do not represent it either truly or falsely:

> 'Tis impossible, therefore, that this passion can be oppos'd by, or be contradictory to truth and reason; since this contradiction consists in the disagreement of ideas, considered as copies, with those objects which they represent (T 415).

Only when a passion is accompanied by a false judgment, either about the existence of an object of the passion, or about the best means for attaining that object, can it be said to be contrary to reason; "and even then 'tis not the passion, properly speaking, which is unreasonable, but the judgment" (T 416). Thus just as reason can no more oppose the passions than a logical argument

could oppose a stone falling through the air, similarly the passions can no more be contrary to reason than a falling stone can be contrary to a logical argument.

This is the context in which one of Hume's most explicit avowals of the negative utility-maximization thesis must be understood. Directly following the argument that a passion can be opposed to reason only in that the judgment which accompanies it might be, Hume says

> (A) 'Tis not contrary to reason to prefer the destruction of the whole world to the scratching of my finger. 'Tis not contrary to reason for me to chuse my total ruin, to prevent the least uneasiness of an *Indian* or person wholly unknown to me. 'Tis as little contrary to reason to prefer even my own acknowledg'd lesser good to my greater, and have a more ardent affection for the former than the latter. A trivial good may, from certain circumstances, produce a desire superior to what arises from the greatest and most valuable enjoyment; nor is there anything more extraordinary in this, than in mechanics to see one pound weight raise up a hundred by the advantage of its situation (T 416).

In passage (A) Hume apparently means to exemplify his previous argument by citing a few illustrations of passions one might think, at first glance, were contrary to reason. But Hume means to provoke us, through these illustrations, into further reflection on his argument, and ultimately into arriving at the opposite conclusion.

This plan is glaringly unsuccessful. First of all, as Baier points out, it would be contrary to reason, in fact self-contradictory, to "prefer the destruction of the whole world to the scratching of my finger," because my finger is part of the world.[18] Second, Hume has just argued that a passion, "such as hope or fear, grief or joy, despair or security" (T 416), cannot be contrary to reason because it "contains not any representative quality, which renders it a copy of any other existence or modification" (T 415). But passions must take intentional objects: We hope *for* something, are afraid *of* something, despair of, over, or about something. Hence the sense in which they contain no "representative quality" is obscure at best.[19]

To be sure, Hume carefully distinguishes between a passion, the cause of the passion, and the object of the passion. A passion, as defined by Hume, is a "violent and sensible emotion of mind, when any good or evil is presented, or

---

[18] Annette Baier, *A Progress of Sentiments: Reflections on Hume's Treatise* (Cambridge: Harvard University, 1991), 165.

[19] Baier rightly describes this as "one very silly paragraph that has perversely dominated the interpretation of [Hume's] moral psychology … [and] is, at the very least, unrepresentative of Hume's claims about passions in the preceding and following parts of Book Two." *Ibid*., 160.

any object, which, by the original formation of our faculties, is fitted to excite an appetite" (T 437). Thus the passion, strictly speaking, is merely the set of physiological and psychological sensations caused by some object or circumstance. In itself, this set does not represent anything; it is an "original modification of existence."

In discussing the indirect passions of pride and humility, Hume also distinguishes between the cause of the passions and their objects:

> … betwixt that idea, which excites them, and that to which they direct their view, when excited. ... The first idea, that is presented to the mind, is that of the cause, or productive principle. This excites the passion, connected with it; and that passion, when excited turns our view to another idea (T 278).

In the context of this discussion, Hume means to distinguish as the cause of the passion that intentional object we feel pride or humility *about*:

> Every valuable quality of mind ... wit, good sense, learning, courage, justice, integrity; all these are the causes of pride. ... A man may [also] be proud of his beauty, strength, agility. ... But this is not all. ... Our country, family, children, relations, riches, houses, gardens, ... any of these may become a cause either of pride or of humility (T 279).

The object of the passion, on the other hand, is in each case the self, i.e. that object in relation to which the ideas of "valuable qualit[ies] of the mind, ... the body likewise, ... [and] whatever objects are in the least ally'd or related to us" (T 279) can excite such sentiments in us.[20] Thus the cause of a passion for Hume is that which we might be inclined to describe as its intentional object, while the object of the passion for Hume is equivalent to what we might describe as its cause, i.e. self-aggrandizement.

However, in discussing the direct passions of desire and aversion, grief and joy, hope and fear, and volition (T 438), Hume often equates the object of a desire, i.e. that to which the passion is directed, with what he calls its cause, for example, when he claims that contrary passions *arise* from different objects of desire or aversion respectively (T 441, 443). Here he allows the possibility that that which causes a passion, e.g. a freshly-baked apple pie, can be the object of the passion as well.

Nevertheless, in spite of Hume's care in distinguishing the cause and intentional object of a passion from the passion itself, it is not plausible to argue that passions cannot be irrational on the ground that in themselves they do not represent or judge anything. For this distinction between the passion and its intentional object is suspect. It is not easy to imagine how we might

[20] Also see Annette Baier, "Hume's Analysis of Pride," *The Journal of Philosophy LXXV*, 1 (January 1978), 27-40.

identify a particular passion independently of its intentional object. Surely we need the death of the close friend, the threat of violence, or the sight of the disgorged calf hanging in the butcher's window in order to distinguish respectively grief, fear, or aversion. The knotting of the stomach, increased heart rate, and tightness in the temples alone do not suffice to distinguish between them – nor, indeed, from particularly intense pleasurable experiences of certain sorts. The intentional object of the passion is part of what identifies it as a particular passion.

Furthermore, it is difficult to imagine a case in which the intentional object of the passion is not a necessary part of the cause of the passion, as Hume rightly suggests. Even neural stimulation would not disconfirm this hypothesis. But these two considerations taken together suggest that a passion always includes, or at least is accompanied by, some "representative quality," i.e. that object which is intentionally represented. So either passions are intrinsically representational, or else they are always "accompany'd with some judgment or opinion" concerning "the existence of objects" (T 416).

This conclusion is borne out by the examples Hume cites in passage (A), all of which make reference to intentional objects. Surely it is at least the ideas of the destruction of the whole world and of the scratching of my finger that causes me to prefer the one to the other; surely it is at least the idea of the unknown Indian that causes me to desire to prevent his uneasiness more than my total ruin. Indeed, it is hard to imagine giving a complete description of any particular passion without referring to its intentionally represented object. But this means that passions can, then, be unreasonable, or contrary to reason after all, for they always involve at least a "supposition of the existence of the object" (T 416), about which one may be mistaken.

Of course a subject need not suppose the object of a passion to have material, empirical existence. Hume would scarcely maintain that any such object must be supposed already to exist in this strong sense. For this would imply that we could only aspire to bring into material existence that which already had it; hence that the desire to achieve or realize our ends played no part in motivating us to action. There is no reason to think Hume held this view. Nor is this supposition required by Hume's notion of intentional existence:

> To reflect on anything simply, and to reflect on it as existent, are nothing different from each other. That idea, when conjoined with the idea of my object, makes no addition to it. Whatever we conceive, we conceive to be existent (T 66-67).

When we conceive of some object or state of affairs we deplore, or wish to attain through action, we suppose it to exist as intentional object of our grief

or desire respectively. We add nothing to this conception by ascertaining whether it exists in a stronger, material sense as well.

But this supposition is a judgment made by our reason, and can be true or false, for it is possible to deplore or desire something that cannot exist even in the weak sense, i.e. a self-contradictory object, such as a department chair who is both just and also partial to oneself. This is the only kind of object which can exist neither as a conceived possible empirical reality to be attained through some course of action, nor as that actual state of affairs which caused the subject to conceive it in the first place. But we might nevertheless mistakenly suppose it could. We might fail to recognize the self-contradictory nature of the object (for example, as when the desired end is to ensure the intact survival of my finger compatibly with the destruction of the whole world). Thus the suggestion is that we understand Hume's criterion of irrationality as involving a mistaken supposition about the intentional existence of the object and not its material existence: We are irrational, in this sense, if we conceive a state of affairs which, because it is internally inconsistent, cannot even be a genuine object of a passion.

This implies that Hume's claims in passage (A) are correct, but not for the reasons he gives. Hume's overall strategy has been to advance a variant on Hutcheson's claim: From the purported non-irrationality of the preferred ends cited in passage (A), we are to conclude the similar imperviousness to rational criteria of our motives. And he has partially succeeded in this enterprise. The choices and preferences he cites are not indeed irrational, but not because they "contain no representative quality" and hence cannot be contrary to reason. They are not irrational because they do not violate the only requirement on ends that Hume by implication proffers: internal logical consistency. But this is in truth no constraint on the range of possible objects of desire at all. It requires merely that any such intentional object be a possible object of desire, i.e. that it not be self-contradictory; and not that it conform to any further requirements such objects themselves must satisfy.

Thus passage (A) provides strong evidence for the negative utility-maximization thesis. For here Hume maintains explicitly the immunity to rational criticism of ends one might intuitively regard as irrational. And he implicitly maintains the conformity of any such end to the requirement that they be possible ends at all. But clearly, this is to require merely that an end be an end. It is not to require that it be rational. Hence, it seems, the corresponding passion is immune to rational criticism as well.

But now we must ask whether Hume's, as well as Hutcheson's, overall argument proves what these writers suppose it proves. Does it in fact follow from the fact that reason imposes no constraints on possible ends that it imposes no constraints on their corresponding passions? The connection

between rational ends and rational motivation is surely not as intimate as Hume and Hutcheson appear to think. For even if we accept the necessary conjunction of a passion with its intentional object, this commits us to the necessary conjunction neither of the passion with its sufficient cause, nor of the passion with any particular end that passion may cause us to desire.

Many things can cause us to feel, say, joy. Remembering something achieved or overcome may cause us to feel joyful. The thing achieved or overcome is then the intentional object of the passion, and also originally causes it. But it can also be the intentional object of the passion without being a sufficient cause of it, as would be the case if it were not the memory of our previous achievements, but rather someone's present praise of them, which causes us to feel joy in those past achievements. Similarly, the feeling of joy in our past achievements may extend into joyful anticipation of future ones. Here the object of the feeling of joy would be a desired end, i.e. anticipated future achievements, while its cause would be the remembered past ones. Thus an identifiable passion – joy in something – is logically independent of both its cause and the end it causes us to desire. Either can function as the intentional object of the passion. Although we require some such intentional object in order to be able to identify the passion, this object need be strictly identifiable with neither its cause nor its desired end.

However, either its cause or its desired end may motivate an agent to action. Joy or pride in our past achievements may move us to take on some new challenge, independently of our enthusiasm for that new project in itself. Or, it may be just and only our enthusiasm for that new project which moves us to action, independently of the feelings of anxiety, fear, uncertainty, or self-doubt it may simultaneously cause us to have. Since a passion can take either its cause or its end as its intentional object, the immunity to reason of its end does not necessarily imply the immunity to reason of that passion itself.

Now suppose it true, as has already been argued, that a passion cannot be unreasonable or irrational, even if it must contain an intentional object. Does this imply that its ends also cannot be unreasonable or irrational? At first glance it would appear that this does not follow. For if the passion can be distinguished from its desired end (as, for example, in the case where the passion's intentional object is its cause but its cause is not its end: My joyful memory of past achievements causes me to take on a new challenge, even though I do not desire that challenge in its own right), then to show that a passion cannot be irrational proves nothing about its end. Apparently, the passion could be immune to rational criticism although its end were not.

But within Hume's framework, this appearance is misleading. For although an end can be detached from *some* passions, such as joy, enthusiasm, grief, or reluctance, it cannot, for Hume, be detached from desire or aversion

for that end. This is the only basis on which Hume permits the object in question to count as an end for us at all (T 414); and desire and aversion themselves are direct passions.

Many states of affairs may cause us to desire something. Among those not identical with the object of the desire are envy, malice, generosity, etc. But in addition to these causes, we must also count as necessary, if not sufficient, the thought of the object itself, considered as a source of pleasure or pain. We cannot experience that passion Hume calls "desire" without simultaneously experiencing the thought of that object our desire is a desire for. So the object of desire, or end, is a necessary concomitant of at least two of the passions: desire and aversion.

Moreover, desire or aversion must be necessary concomitants of all the other passions, for Hume, in so far as these motivate the agent to seek an object of pleasure or avoid an object of pain (T 414, 417). We could not blindly take on the new project, merely out of joy in our past achievements. For this alone would not be sufficient to determine our choice of that one end over many others. Out of joy in our past achievements alone we might as easily choose to rest on our laurels as to press on to something new. Although this joy might well override any fondness or enthusiasm we might feel for the end in its own right, there must be at least enough interest to determine our choice of that end rather than some other; and Hume supplies no alternative to desire, for example an account of intention as causally efficacious, that would satisfy this desideratum.[21]

Hence for Hume the very fact that we adopt some particular end indicates the presence of a desire for that end. Conversely, the presence of desire is sufficient to indicate an end or purpose, since desire is one of those passions that must take an intentional object. Hence the presence of desire can be tautologically construed as a necessary ingredient in any combination of passions that can motivate us to action, just as the contemporary belief-desire model of motivation would require. So if the passions are the sole sources of behavioral motivation, and if the passions cannot be contrary to reason, then the ends they lead us to adopt cannot be irrational either. The absence of rational constraints on desire is both a necessary and a sufficient condition for the absence of rational constraints on ends. Thus we must conclude that Hume not only accepts the traditional view of reason, but actively embraces both the positive and the negative utility-maximization theses – for more reasons even than he himself explicitly gives.

[21] As, for example, Kant arguably does.

4. The Principles of Variability

I now consider an argument that may incline some Hume scholars to an opposite conclusion, i.e. that in spite of the evidence to the contrary already assembled, Hume does in fact provide a positive account of what amounts to rational constraints on ends.[22] In Book I, Part IV, Section 4 of the *Treatise*, Hume distinguishes between those principles

> which are permanent, irresistible, and universal; such as the customary transition from causes to effects, and from effects to causes: And the principles, which are changeable, weak, and irregular. ... (T 225)

such as the superstitious inclination to impute a faculty or occult quality to phenomena we cannot otherwise explain (T 224). He argues that the former are received by philosophy for the simple reason that human life would be impossible without them: "[They] are the foundation of all our thoughts and actions, so that upon their removal human nature must immediately perish and go to ruin" (T 225). Of course Hume does not claim that such "permanent, irresistible, and universal" principles – let us call them *PIU principles* – are rational, nor that they are logically or conceptually necessary. They are necessary merely for the survival of human nature, of our capacities for thought and action. But as I have argued in Chapter XIII.7, philosophy is, even for Hume, the discipline of rational thought *par excellence*. So it might be argued, at least, that the reception of the PIU principles by philosophy is strong *evidence* of their rationality.

Later, in discussing the problem of freedom of the will in Book II, Hume identifies those natural principles which govern human behavior as being of a piece with PIU principles. He argues, for example, that

> whether we consider mankind according to the differences of sexes, ages, governments, conditions, or methods of education; the same uniformity and regular operation of natural principles are discernible. Like causes produce like effects; in the same manner as in the mutual action of the elements as powers of nature (T 401).

Hume then goes on to assert that just as the cohesiveness of matter arises from necessary principles, similarly, human society is founded on principles that are just as necessary. Indeed, we can be even more certain of such necessary natural principles governing human social phenomena than we can in the

[22] The argument as I present it is a variant on that offered by David Miller, 37-39 (*op. cit.* Note 2), although Miller does not claim rational, but rather merely reflective and analytical status for the PIU principles. I am grateful to Louis Loeb for originally calling my attention to the passages on the PIU principles, and for discussion of them, although the use I make of them here is my own. Loeb develops this notion in a different direction in "Cartesian Epistemology Without Divine Validation of the Cognitive Faculties" (unpublished paper, 1985).

case of natural phenomena, for we are more successful in explaining the former than the latter:

> [T]he different stations of life influence the whole fabric, external and internal; and these different stations arise necessarily, because uniformly, from the necessary and uniform principles of human nature. ... There is a general course of nature in human actions, as well as in the operations of the sun and the climate. There are also characters peculiar to different nations and particular persons, as well as common to mankind. The knowledge of these characters is founded on the observation of an uniformity in the actions, that flow from them; and this uniformity forms the very essence of necessity (T 402-3).

What are the certain principles of human behavior that Hume has in mind? These can be divided into two categories: (1) those principles describing the influence of sensory limitations and the violent passions on human behavior, which I shall refer to as *principles of variability*; and (2) those describing the modifying influence of the calm passions, which I shall call *principles of stability*.[23] A violent passion is, as we saw, a "violent and sensible emotion of mind, when any good or evil is presented" (T 437), whereas calm passions are "affections of the very same kind ... but such as operate more calmly. ..." (T 437)

> tho' they be real passions, produce little emotion in the mind, are more known by their effects than by the immediate feeling or sensation. These desires are of two kinds; either certain instincts originally implanted in our natures, such as benevolence and resentment, the love of life, and kindness to children, or the general appetite to good and aversion to evil, consider'd merely as such (T 417).

Whether a passion is calm or violent depends on the individual's temper, the circumstances and situation of the object, the intensity of other simultaneous passions, its degree of habituation, and the extent to which it excites the imagination (T 438).

Hume's account of the relationship between (1) and (2) is basically as follows. Possible objects of desire undergo modification and distortion in perceived degrees of desirability, accordingly as the passions that adopt them vary in violence or intensity (or "vivacity"), and as other contingent conditions vary. The variability in the violence of the passions depends upon just the

[23]By contrast, Miller (*ibid*.) takes Hume's PIU principles to refer solely to general, higher-order rules by which our first-order beliefs and inferences can be corrected (see T 146-50, and Book I, Part III, Section 15, "Rules by which to judge of causes and effects"). This is where my understanding of the PIU principles diverges from Miller's: Miller thinks Hume means to refer only to principles governing our judgments, whereas I contend that he means to refer to principles governing our behavior more generally.

contingent circumstances that generate them. However, the distortive effect of these circumstances is partially corrected by the operations of the calm passions, which are often mistaken for reason. Let us now examine this account more closely. I treat Hume's principles of variability in this section, leaving his principles of stability for Section 5. Finally, in Section 6, I again recur to and dispose of the general argument that claims that Hume does, in effect, impose rational constraints on ends.

In the *Treatise*, Hume enumerates the principles falling into the first category in greater detail:

> (a) We are more inclined to pursue a good when it is near to us than when it is remote, because the nearer it is the more violent the passion it causes, and we are more easily impelled to action by violent than by calm passions (T 319; also 427-34).
>
> (b) Similarly, we are more strongly impelled to pursue or avoid an object about which we experience conflicting passions than we would be otherwise, for these increase the intensity of the predominating passion we feel toward it (T 421).
>
> (c) Uncertainty in the apprehension or prospects of realizing the object, on the other hand, tends to increase our enthusiasm for it much as security tends to replace enthusiasm with boredom (T 421-22).
>
> (d) Custom and repetition in the performance of certain actions can transform the accompanying violent passion into a calm one. For they give rise to a facility in performing the action. On the one hand, this facility is an additional source of pleasure (up to a certain point) that motivates us to repeat the action. On the other hand, repetition transforms the action into a settled habit of conduct we perform without feeling intensely motivated to do so (T 422-4; cf. 426).
>
> (e) Finally, our imagination increases our pleasurable anticipation of achieving some object, insofar as our prior experience of it enhances our conception of it, as does our memory of it (T 424-6).

These are most prominent among Hume's principles of variability. In a significant passage in the *Enquiry*, to which I shall recur, Hume summarizes these circumstances when he maintains that

> when some of these objects approach nearer to us, or acquire the advantages of favorable lights and positions, which catch the heart or imagination; our general resolutions are frequently confounded, a small enjoyment preferred, and lasting shame and sorrow entailed upon us (E 239; cf. T 536).

Is it only objects of desire that we must appraise cautiously in order to correct our distorted or prejudiced perceptions of them? Are objects of desire the only subjects of principles of variability? Hume has already answered this question in the negative. It is not merely the violence of our passions that color our perceptions, but our sensory limitations as well:

> [T]he senses alone are not implicitly to be depended on; ... we must correct their evidence by reason, and by considerations, derived from the nature of the medium, the distance of the object, and the disposition of the organ, in order to render them within their sphere, the proper criteria of truth and falsehood (E 117/151).

Thus *all* objects of perception, including objects of desire, are subject to the distortions arising from the limitations of individual circumstances: our spatiotemporal relation to the object, our personal constitution, the psychological background against which we apprehend the object, and the intensity of the sentiments aroused by it.

Hume makes equally clear that it is not only perceived objects that are susceptible to this distortion, but perceived subjects as well. In the *Enquiry*, Hume argues eloquently that all human beings have instincts of sympathy and benevolence, even if these vary enormously among individuals and circumstances. Two factors determining the intensity or violence of our sentiment of sympathy or approval for someone's moral behavior are (1) the extent to which the person's actions affect us personally; (2) the person's spatiotemporal proximity to us. Hence our sentiments are more deeply aroused by a statesman serving our own country, now, than by one serving another country or one whose actions occurred in the distant past (E 185/227). Hume explicitly maintains that we must correct the inequality of our responses to the two cases in the same way, and for just the same reasons, as we must when making perceptual judgments or choosing among desired objects:

> [W]here the good, ... [is] less connected with us, [it] seems more obscure, and affects us with a less lively sympathy. We may own the merit to be equally great, through our sentiments are not raised to an equal height, in both cases. The judgment here corrects the inequalities of our internal emotions and perceptions; in like manner, as it preserves us from error, in the several variations of images, presented to our external senses. ... And, indeed, without such a correction of appearances, both in internal and external sentiment, men could never think or talk steadily on any subject; while their fluctuating situations produce a continual variation on objects, and throw them into such different and contrary lights and positions (E 185/227-8).

This argument is derived, in essence, from a similar one Hume makes in the *Treatise*. There he is concerned to refute the objection that since our moral sentiments vary while our moral appraisals do not, these appraisals are not based on our moral feelings but rather on reason. Hume's response is that our moral judgments themselves are based on "a moral taste, and from certain sentiments of pleasure or disgust, which arise upon the contemplation and view of particular qualities or characters" (T 581). Hume's point here is an important one: It is that judgments, thought to issue from reason conceived as distinct from the passions, are not in fact independent of those passions or sentiments, but rather are generated by them. Thus the same feelings – pleasure or aversion – arise in response to perceiving moral qualities as they do in response to other sorts of possible objects of desire. Hume concedes, as before, that these sentiments

> must vary according to the distance or contiguity of the objects ... our situation, with regard both to persons and things, is in continual fluctuation. ... Besides, every particular man has a peculiar position with regard to others; and 'tis impossible we cou'd ever converse together on any reasonable terms, were each of us to consider characters and persons, only as they appear from his particular point of view (T 581).

So moral judgments about persons as well as nonmoral ones about objects of desire and perception are susceptible to distortion, insofar as they are colored by our own variable circumstances. Each of these types of objects contribute to the subject matter of Hume's principles of variability, for each is a type of object with respect to which our judgment must be distorted by the very subjectivity of our situation itself. I shall call the perception conditioned by this situation the *subjective perspective*.

I suggested that the calm passions are claimed by Hume to provide a partial corrective to the subjective perspective, and that their workings constitute the subject matter of what I termed *principles of stability*. In the following section, I elaborate this suggestion in detail.

5. The Principles of Stability and the Objective Perspective

Hume immediately continues the above discussion by arguing that we correct these variations in our sentiments and perceptions by fixing on what he describes as "some *steady* and *general* points of view; and always, in our thoughts, place ourselves in them, whatever may be our present situation" (T 581-2). He contrasts this steady and general point of view with the actual variations in viewpoint that occur because of the changes in our particular circumstances, arguing that our use of language disregards such fluctuations, and expresses "our liking or dislike, in the same manner, as if we remained in one point of view" (T 582). However, he contends, we do not thereby fully

succeed in correcting the waywardness and partiality of our feelings through behavior that is consistent with this stable and general view:

> [R]eason requires such an impartial conduct, but ... 'tis seldom we can bring ourselves to it, and ... our passions do not readily follow the determination of our judgment. This language will be easily understood, if we consider what we formerly said concerning that *reason*, which is able to oppose our passion; and which we have found to be nothing but a general calm determination of the passions, founded on some distant view or reflexion (T 583).

The last sentence summarizes Hume's earlier argument of Book II, Part III, that reason, far from opposing and controlling the passions in the service of morally obligatory behavior, is in fact of a piece with them, and that we mistake certain passions for the motivating influence of reason only because they operate tranquilly rather than violently on us (T 417, 437).

But for our present purposes, this passage is significant for the additional light it sheds on the "steady and general view" that corrects the contingencies of our individual perspectives. For here Hume further characterizes this view as impartial, reflective, distant, often mistaken for the operations of reason, and the basis for a "general calm determination of the passions." Thus the basic picture is that of a perspective that corrects for individual contingencies, changes, and partiality of vision by being stable where individual perception is fluctuating; general where individual perception is confined to the particular perspective dictated by its own relation to the object; impartial or judicious where individual perception is biased in its view by its location relative to the object; and reflective where individual perception is impulsive and unselfconscious in its appraisal of the object. Finally, this perspective provides the foundation for the tranquil and undisturbed workings of the calm passions, which are consequently mistaken for the operations of reason. Let us call this the *objective perspective*.

The basic argument in support of the objective perspective would appear to be as follows:

> (P.1) Nearness and remoteness to the object of appraisal is a function of psychological as well as spatial or temporal proximity to the individual;
>
> (P.2) The violence and intensity of our passions decrease with the object's psychological distance from the self, much as they do with its spatial or temporal distance from the physical location of the individual;
>
> (C) The greater the spatiotemporal or psychological distance of the object from the individual, the more nearly we approach the objective perspective.

That Hume maintains (P.1) follows from the variety of objects he subjects to his principles of variability, of which we have already spoken. (P.2) follows from his many and detailed discussions of the disturbing and distinctive effects of the object's spatiotemporal and psychological proximity to the individual, which we have also already reviewed (e.g. T 489, E 234). (C) follows from the premises plus the implicit assumption that the objective perspective is to distance as the subjective perspective is to proximity. We find support for this assumption in Hume's own repeated use of the phrase "distant view" to characterize this perspective (e.g. T 583, E 196/239). We can then further describe the objective perspective as one that involves psychological and emotional distance from just those objects that are psychologically and spatiotemporally – therefore emotionally – closest to us: considerations of self-interest, immediate sources of pleasure, proximate objects of gratification, etc. To distance ourselves from these objects is precisely to view them as though from that psychological or spatiotemporal distance at which they would not affect the passions as violently and distort our judgment as completely as they otherwise do.

This interpretation is further confirmed by the following important passage from the *Enquiry*, which I quote in full:

> (B) All men, it is allowed, are equally desirous of happiness; but few are successful in the pursuit; one considerable cause is the want of strength of mind, which might enable them to resist the temptation of present ease or pleasure, and carry them forward in the search of more distant profit and enjoyment. Our affections, on a general prospect of their objects, form certain rules of conduct, and certain measures of preference of one above another: and these decisions, though really the result of our calm passions and propensities (for what else can pronounce any object eligible or the contrary?) are yet said, by a natural abuse of terms, to be the determinations of pure *reason* and reflection. But when some of these objects approach nearer to us, or acquire the advantage of favorable lights and positions, which catch the heart and imagination; our general resolutions are frequently confounded, a small enjoyment preferred, and lasting shame and sorrow entailed upon us. And however poets may employ their wit and eloquence, in celebrating present pleasure, and rejecting all distant views to fame, health, or fortune; it is obvious that this practice is the source of all dissoluteness and disorder, repentance and misery (E 196/239).

In passage (B) Hume makes a number of important points. First, he amplifies further his conception of the objective perspective. For here we see that this perspective requires us not merely to distance ourselves emotionally from our

most proximate interests, objects of desires, and appraisals, but explicitly to assume the vantage point of a psychologically or spatiotemporally remote interest, object of desire, or appraisal in order to achieve this.

These two are distinct. I can detach myself from my closest concerns by emotionally withdrawing from them. By repressing diminishing, or subduing the intensity of my desire for a Black Forest Torte, I achieve a certain detachment from this desire. It ceases to upset my composure, hence permits me to reflect on it more tranquilly, or consider with greater liberality features of it that my emotional investment in it might otherwise obscure or bypass altogether. A person who is not temperamentally susceptible to tempestuous feelings is able to view most of his interests and desires with greater intellectual clarity and equanimity, for it allows him to analyze and explain such things without the unbalancing impediment of emotional involvement.

But emotional detachment is not sufficient for achieving the objective perspective. For it does not follow from my lack of emotional upheaval over my most proximate objects of desire or appraisal that I therefore do not, because of their proximity, mistakenly ascribe to them primary value. That is, it does not immediately follow from the assumption that the calm passions are governing one's behavior that one thereby appraises objects of desire judiciously. It is hardly unusual to encounter a person who is both calm and biased; whose emotional tranquility is matched only by a staunch conviction in the primacy of her personal interests above general ones. Hence it is not enough to distance oneself merely from the distorting effects of the violent passions, for this degree of detachment is nevertheless consistent with maintaining the subjective perspective. Unbiased and judicious judgment requires, in addition, that *one view one's subjective perspective itself* from a distance. And this requires not just emotional detachment, but intellectual and psychological distance from one's concerns as well. Hume's specification that one assume the vantage point of distant concerns makes this requirement explicit.

However, concerns can be distant in two ways. They can be distant from the constellation of interests, desires, beliefs, and judgments that constitute my present self, but nevertheless proximate to the constellation that I now know will compromise my future self, or my overall self considered through each moment of time. This would be the stance of enlightened self-interest. Alternately, concerns can be distant from my self *simpliciter*, i.e. such as will never constitute part of myself from any temporal perspective, hence can never be subsumed under the rubric of self-interest. This would be the transpersonal stance of *strict impartiality*. I discuss the concept of impartiality in detail in Volume II, Chapter VI.

Some have interpreted Hume's sketchy remarks about the objective perspective, both in the *Treatise* and in the *Enquiry*, as referring to the stance of strict impartiality.[24] And indeed this interpretation is supported by Hume's claim in the *Treatise* that

> 'Tis seldom men heartily love *what lies at a distance from them, and what no way redounds to their particular benefit*; as 'tis no less rare to meet with persons, who can pardon another any opposition he makes to their interest, however justifiable that interest may be by the general rules of morality. Here we are contented with saying, that reason requires such an *impartial conduct* but that 'tis seldom we can bring ourselves to it: and that our passions do not readily follow the determinations of our judgment (T 583; emphasis added).

On this reading, the objective perspective is mistakenly thought to be equivalent to the perspective of reason, which dictates objectively and impartially without regard for the claims of self-interest. The difficulty is that it is not immediately clear how this perspective is to be achieved by any limited sentient individual, nor how it is even connected with the subjective perspective with which every individual is familiar.

Passage (B) from the *Enquiry* indicates that it is rather the stance of enlightened self-interest that Hume has in mind. There Hume is fulminating against the evils and misery of pure time preference, i.e. of preferring some satisfaction over another purely because of its greater temporal proximity to the agent. He is recommending that we detach ourselves from the satisfactions of the immediate present, and choose objects or courses of action with a view to our future happiness, or our happiness considered as a whole, over the entire course of our lives. We are to think of our overall, genuine rather than our immediate self-interest. But it is a far cry from this distance from some one time-slice of my life to the greater, quite dizzying distance from all time-slices of all lives that is necessary for judging any one such time-slice from the transpersonal stance of strict impartiality.[25] For in the *Treatise*,

---

[24] For example, Stephen Darwall (*Impartial Reason* (Ithaca, New York: Cornell University Press, 1983), 60) takes Hume to be committed to this brand of distance when he maintains in the *Treatise* that "'Tis only when a character is considered *in general without reference to our particular interest*, that it causes such a feeling or sentiment, as denominates it morally good or evil." (T 472; emphasis added) Also see Marcia Baron, "Hume's Calm Passions," (M. A. Thesis, The University of North Carolina at Chapel Hill, 1978).

[25] In Chapter VII I have discussed Thomas Nagel's early analysis of this concept in *The Possibility of Altruism* (New York: Oxford University Press, 1970). But a more refined account that raises correspondingly more issues is to be found in his "Subjective and Objective," in *Mortal Questions* (New York: Cambridge University Press, 1979), 196-213.

Hume takes impartial judgment to be the opposite of self-interested judgment of any kind. Strictly impartial judgment then requires a distant view that is nevertheless not the view of any one *self* at all, neither immediate nor future, nor unified as a whole over time. It is difficult to say, within the Western Anglo-American analytic tradition, in what such a view might consist.

There is thus good reason why Hume may have opted, upon mature reflection, for the stance of enlightened self-interest described in the *Enquiry*. To be sure, it rules out strict impartiality by presupposing that our distant view is nevertheless always the view of one's self, hence that its appraisals of objects are conditioned accordingly. But it simultaneously makes room for a more limited, intermediate distance that at the same time satisfies the requirement of the objective perspective, i.e. that we transcend the distortions contingent on considerations of immediate self-interest to achieve judiciousness in our judgments. We find an account of this intermediate distance, and how it is achieved, in Hume's claim that

> [e]very man's interest is peculiar to himself, and the aversions and desires, which result from it, cannot be supposed to affect others in a like degree. General language, therefore, being formed for general use, must be moulded on some more general views, and must affix the epithets of praise or blame, in conformity to sentiments, which arise from the general interest of the community. ... Sympathy, we shall allow, is much fainter than our concern for ourselves, and sympathy with persons remote from us much fainter than that with persons near and contiguous; but for this very reason it is necessary for us, in our calm judgments and discourse concerning the characters of men, to neglect these differences, and render our sentiments more public and social. Besides, that we ourselves often change our situation in this particular, we every day meet with persons who are in a situation different from us, and who could never converse with us were we to remain constantly in that position and point of view, which is peculiar to ourselves. The intercourse of sentiments, therefore, in society and conversation, makes us form some general unalterable standard, by which we may approve or disapprove of character and manners (E 186/228-9).

This passage enumerates three steps that permit us to move from the subjective to the objective perspective:

Many of these are resolved in his yet more recent *The View From Nowhere* (New York: Oxford University Press, 1985).

(1) We discount the characteristics that distinguish between ourselves and others, and between persons near to us and those remote from us;

(2) We note consciously the "intercourse of sentiments" consequent on our regularly and often changing our own positions and exchanging it for those of others in society with whom we must communicate;

(3) We form a more generalized conception of the features common to both of our situations.

(1) enables us to overcome the limitations of our individual vantage points. (1) by itself, however, would not suffice for the detached, objective perspective, for it would, as already pointed out, leave us with no particular point of view at all from which to regard them. (2) then stipulates that alternate point of view: that of the other individuals collectively, with whom we interact – Rawls's and Habermas's "we-perspective." By putting ourselves in these other situations, we gradually develop from a subjective, enclosed view of our concerns a more general one that encompasses the common features of all the perspectives of those with whom we have exchanged positions and sentiments, just as Habermas in particular recommends. This is step (3), the "general, unalterable standard" by which we then make normative judgments and which arise from the general interests of the community. Thus the "general interests" are those which remain invariant across exchange of positions and sentiments among individuals. Clearly these must include certain of the self-interests of any one of these individuals chosen at random.

Further evidence for this reading of the objective perspective as the stance of enlightened self-interest can be culled from Hume's discussion of the "common interest" in his treatment of justice and property in the *Treatise*. In discussing the origin of the convention to respect private property, he says of it,

> It is only a general sense of common interest; which sense all the members of the society express to one another ... I observe, that it will be for my interest to leave another in the possession of his goods, *provided* he will act in the same manner with regard to me. He is sensible of a like interest in the regulation of his conduct. When this common sense of interest is mutually express'd, and is known to both, it produces a suitable resolution and behavior ... the sense of interest has become common to all our fellows, and gives us a confidence of the future regularity of their conduct. ... In like manner do gold and silver become the common measures of exchange, etc. (T 490; emphasis in text).

This description of how the conventions of private property, language, and money are established satisfies the three-step sequence for moving from the subjective to the objective perspective of the common or general interest, and thereby supports Hume's remarks in the *Enquiry* about the relation of language to the general interests of the community: I begin by observing the differences between my own position (as possessor of some good) and that of the other (as potential threat to my possession). I then discount those differences (step (1)). Next, I exchange our respective positions: He, as possessor of goods, is as much threatened by my potential aggression as I was by his (step (2)). In step (3), we each recognize our common features as possessors of goods with an interest in protecting it, and it is the recognition of this common interest that then establishes the convention of conduct, i.e. respect for private property, which allows each of us to satisfy it. The same reasoning can be applied to the conventions of language or money.

The general point is clear: Establishing the social conventions that make human society of any kind whatsoever possible requires moving from a narrow, subjective view of our own interests that distorts our appraisal of different states of affairs to a more objective perspective that regards those interests from the viewpoint of the interests shared by the community as a whole (E 186/229, fn.). This objective perspective enables one to appraise some state of affairs, but not with strict impartiality; for I have suggested that this is in any case metaphysically impossible on Hume's view. Rather, it enables us to appraise it judiciously, in the sense that we can view the matter from the vantage point of the community's interests. And it is only this perspective that allows us to establish the conventions of behavior on which human society can be erected.

These remarks illuminate the second important point Hume makes in passage (B), i.e. that the subjective perspective is the "source of all dissoluteness and disorder, repentance and misery." The greater the uncorrected proximity of the objects of desire, the more we are victimized by the violent passions they produce, and the more unconsidered and disorderly are our actions in their pursuit. The subjective perspective is, then, the source of moral and personal chaos that undermines social order and the conventions that maintain it. It is a threat to the general interest that the objective perspective so clearly recognizes:

> 'Tis certain, that self-love, when it acts at its liberty ... is the source of all injustice and violence; nor can a man ever correct those vices, without correcting and restraining the *natural* movements of that appetite (T 480).

Now we are in a better position to see how the calm passions provide a partial corrective to the subjective perspective. The social conventions that arise out of that recognition of the common or general interest that

characterizes the objective perspective are precisely those actions motivated by calm passions; Hume is quite explicit about this, not only in passage (B), but also in the *Treatise*, where he describe a calm passion as one which "has become a settled principle of action" to which "repeated custom and its own force have made everything yield" (T 419).[26] The calm passions are those which motivate us to perform those habitual and customary actions, or conventions, in which most of social life consists. These are the "certain rules of conduct" formed by our affections on "a general prospect of their objects," and often mistakenly identified as the workings of reason. Passions originally became calm through repetition of the actions they motivate. Thus they are mistaken for the operations of reason, not only because they fail to disturb us emotionally, but also because they result in *general rules* of conduct under which repeated instances are subsumed.

But it is in fact not reason that enjoins us to this customary conduct. Rather, it is the recognition of our genuine self-interest, i.e. of the ends we most desire to achieve. We repeatedly perform those actions because we recognize them as convention solutions to a coordination problem, i.e. how to behave so that common interests are maximized and individual interests are promoted. This is a problem because acting solely in the pursuit of immediate individual interest is to act from the subjective perspective, hence to be victimized by distorted and biased appraisals of where our genuine best interests actually lie. This bias is corrected by that recognition of the common interest that occurs as we move through the three-step sequence into the objective perspective. This recognition in turn enables us to formulate and act upon those rules of conduct that, "when [thus] coordinated by reflection and seconded by resolution, are able to control [the violent passions] in their most furious movements" (T 437-8), hence preserve the social order.

Now those who contend that Hume's introduction of what I have called the objective perspective commits him to ascribing a larger role to reason in motivating action than his explicit arguments suggest, may contend that this "steady, distant, reflective view" on which the workings of the calm passions are founded is not itself a passion but rather a function of the understanding, or reason. I see no reason to accept this contention. The analysis given in these pages suggests that the objective perspective is nothing more than a perception of others' interests, coupled with an absence of those emotional obstacles that usually prevent our recognizing the extent to which those interests coincide with our own. This absence of emotional obstacles does not imply the presence of intellectual cogitation, but rather the presence of

[26]This is consistent with interpreting the prevalence of the calm over the violent passions as a natural virtue (T 418).

tranquil passions lulled into quiescence by repetition and habit. And we win recognition of our common interests not through rational reflection, but rather through having had many and varied social interests with others with whom we do, in fact, have much in common.

Thus the subject matter of what I have termed Hume's principles of stability are those actual rules of conduct in which the calm passions find expression, and which act as an antidote to the disruptive and distorting effects of the violent passions that normally characterize the subjective perspective. These principles are directly antithetical to the principles of variability, in that the latter enumerate the psychological laws by which social order is disrupted through the stimulation of the violent passions, while the former, if spelled out, enumerate the social rules by which it can be maintained.

This completes our discussion of Hume's principles of stability. In closing, it remains only to be reemphasized that for Hume, both principles of variability and principles of stability are uniform and necessary laws of human nature, for they are subject in exactly the same way to the causal determinants that condition any natural event. They are explicitly stated by Hume to be of a piece with – indeed, instances of – the operations of causal law. Next I fit this account into the argument that claims Hume to have in effect imposed rational constraints on ends.

6. The Rationality of Final Ends

According to the argument introduced in Section 4, that these principles of stability are of a piece with causal law implies that they, too, are PIU principles that must be received by philosophy. Now Hume may not explicitly identify these principles as rational. In fact, we have seen that he repeatedly and explicitly denies rational status to the principles of stability. But perhaps these passages are to be collectively discounted, if it can be shown that Hume's principles governing the passions in fact satisfy all the conditions that rational principles must satisfy. For recall Hume's characterization of reason. He distinguished it into demonstrative and probabilistic. And his arguments regarding the status of causal connection, together with his taxonomical division of the faculties of reason, implied that the concern of probabilistic reasoning is causal connection. We now discover that the two kinds of principles describing the operations of the passions are a species of causal law. The inference is evident: The principles governing the passions conform to probabilistic rationality. And to the extent that "our actions have a constant union with our motives, tempers, and circumstances" (T 401), the ends they determine will be equally settled, uniform, and regular. Indeed, this inference finds confirmation on page 281 of the *Treatise*, where

Hume argues, first, that the same objects – power, riches, beauty, personal merit – give rise to the same passions in all nations; and second, that new objects adapt themselves to an already existing passion by partaking of some general quality shared by its other objects, to which the mind is already disposed. Hence the rational principles describing the ways in which the passions typically operate provide an equally rational set of constraints on the ends or intentional objects those passions typically take. It would seem that the PIU principles of the passions do provide a positive set of constraints on the range of ends it is rational for a human agent to adopt.

But this conclusion is mistaken. What is rational about the PIU principles of the passions, if anything, is the fact that they are, like other causal law, necessary, uniform, and general in their application. Moreover, like other causal law, they describe law-like and seemingly regular and predictable relations among given phenomena. It is the fact that they qualify as genuine principles which entitles us to think of them as rational. Similarly, for Hume, it is a certain kind of relation between abstract ideas that is rational, i.e. the inferentially correct and real one. In both cases, we are exercising our reason in so far as we investigate and determine the true – which is to say the uniform, universally valid, and "necessary" connections among given states of affairs.

One may want to argue that Hume's principles of stability are rational in a further sense as well: As effective social rules and conventions, they are rational means to the achievement of individual ends, in that they are the most efficient ways of achieving various states of affairs desired by individuals, consistent with satisfying the common interest in social order. This argument can be illustrated by Hume's treatments of the origin of justice and private property discussed above (respectively, Sections 2 and 5).

But in neither case can this be thought to imply that the states of affairs themselves to which the PIU principles apply are rational. That there is a logical and rational relation between the idea of being a bachelor and the idea of being an unmarried man does not suggest that either idea as such is rational. That there is a causal and probabilistically rational relation between the color of litmus paper and the acid solution in which it is dipped suggests the rationality neither of the color of the litmus paper nor of the relevant solution. And that there is a similar type of relation between the intensity of one's craving for a Black Forest Torte and its actual proximity, or between one's desire to retain one's own possessions and one's respect for those of others, suggests the rationality neither of the craving nor of the Torte nor of private property.

The general point is clear: That there is a rationally discernible relation between the passions and the ends they try to achieve does not imply

the rationality of those ends any more than its does the rationality of the passions themselves. Hence Hume's principles of variability and stability do not delimit a range of identifiably rational ends. For the demand for identifiably rational final ends is not for principles governing ends that are rational in virtue of the rational status of the *principles*. The demand is for principles governing ends that confer rational status on the *ends*. The PIU principles of the violent and calm passions do not meet this demand.

This conclusion follows, indeed, from Hume's very characterization of the passions:

> [W]hat we commonly understand by *passion* is a violent and sensible emotion of mind, when any good or evil is presented, or any object, which, by the original formation of our faculties, is fitted to excite an appetite (T 437).

Hume's first point here is that *when* any object that is good, evil, or capable of causing in us a desire or aversion for it is presented to us, we *then* experience a "violent and sensible emotion of mind," or at least a more tranquil one that "cause[s] no disorder in the temper." His second point is that the range of objects capable of affecting us in this way is constrained only by our own capacity to so respond to it, i.e. by "the original formation of our faculties."

Two implications of Hume's claims follow immediately: First, the passions, both violent and calm, depend on the prior presentation of some object in order to be aroused. It is only if we are already conscious of the object as desirable or repellent that we are then incited to pursue or avoid it. Hence the passion follows rather than precedes adoption of the object as a positive or negative end. This summarizes and is underscored by Hume's earlier assertion that

> 'Tis from the prospects of pain or pleasure that the aversion or propensity arises toward any object: And ... these emotions extend themselves to the causes and effects of that object, as they are pointed out to us by reason and experience (T 414).

This passage occurs as part of Hume's argument that reason can provide no motivation to action. But the temporal priority of perceiving the object as a source of pleasure or pain over the excitation of a motivating passion for or against it stands nevertheless. If we must perceive the object as desirable or undesirable before we are motivated to achieve or avoid it, then it must be a recognizable end for us, whether positive or negative, before we are moved to action on its behalf. But if the recognition of the object as a desirable end is presupposed by its exciting a violent or calm passion, it is not easy to see how the passions might originally determine any particular range of ends. Clearly, it would seem to be the other way around.

The second consequence of Hume's claim, and the conclusion of this chapter, is that on Hume's own view as well as the Humean one, the only constraint on the range of objects that can be possible ends or objects of desire for us is our own motivational capacity. The Humean conception of the self permits us to adopt anything as an end that we can be moved to attain. This diminishes even further the plausibility of supposing that either the passions or the PIU principles that govern them might impose rational constraints on final ends. For Hume as well as for Humeans, such constraints can consist only in our natural capacity for desiring. And the counterintuitive examples of rationally permitted final ends enumerated throughout this volume – counting blades of grass, howling at the moon, and the like – strongly suggest that to argue for the rationality of this capacity as a rational constraint on what we can desire is implausible at best. Hence when Hume flamboyantly but categorically denies that reason can influence our final ends, we must take him at his word, with all the counterintuitive and methodologically exasperating implications to the itemization of which this volume has been devoted. The challenge is then to develop an alternative conception of the self, of motivation, and of rationality that avoid these implications. This is the positive, substantive challenge that I try to meet in Volume II.

Chapter XV. Seven Dogmas of Humeanism

In this volume I have reviewed and critiqued the evolution of the Humean conception of the self in mid- to late-twentieth century Anglo-American moral philosophy through close attention to its use in the hands of several of its leading proponents, as they have developed its foundational notion of desire in response to certain basic dilemmas this conception generates. I have tried to track the ways in which the notion of desire has proliferated from the commonsense, prereflective concept of a desire, to that of desire as a theoretical construct, to that of desire as a dispositional response, to that of an unconscious desire, to that of a behaviorally revealed desire, to that of an internally coherent system of desires, to that of cardinally and then ordinally ranked desires, to the distinction between motivated and unmotivated desires, to that between first- and higher-order desires, to that between self-directed and other-directed desires, to that between object-dependent, principle-dependent, and conception-dependent desires, to that between blindfolded and fully informed desires. And I have tried to show that none of these sophisticated epicyclic refinements of the fundamental notion of a desire solve or avoid the basic dilemmas this notion engenders. I have tried to suggest that no such future epicyclic variations can; that their solution requires a paradigm shift away from desire altogether, and toward reason as the primary foundational factor in both thought and action.

In Volume II I try to show that solutions to the three problems listed in Chapter I.7.2.2 above, in addition to several others left so far untended, require dismantling and reconstructing within a broader, Kantian framework a constellation of familiar, reductive metaphysical dogmas, inherited from Positivism, which the Humean conception presupposes virtually without question. The influence of these dogmas extends far past the confines of Humean moral philosophy. Humean moral philosophy rather takes its cue from these more widespread, Positivist metaphysical doctrines that came to define late-twentieth century Anglo-American analytic philosophy in general. I find a great deal of significance in these doctrines – as I do in Humean moral philosophy itself; and so have no interest in doing away with them. However, it is in the nature of reductive doctrines of any kind to be inherently exclusive of views, assumptions, data and strategies that interfere with the reduction; and if there is one overarching theme of this project, it is that doctrinal and conceptual exclusion is exactly the wrong direction in which twenty-first century Anglo-American analytic moral philosophy ought to be moving. So my aim in Volume II is the less bellicose one of targeting, tempering, and situating the constructive kernels of these dogmas within a larger context of assumptions with which, shorn of their reductive and exclusionary aspirations, they might more or less peacefully co-exist. To that end, I close

this first, critical half of the discussion with a brief conceptual map that locates the Humean conception of the self within its network of Positivist metaphysical dogmas – dogmas which lend one another mutual support and enhance the *prima facie* credibility of the Humean conception in relation to them. And I offer a very brief preview of how I intend to temper them in the second, substantive half of this discussion.

One such dogma that provides a rationale for the Humean position may be found in the *epiphenomenalist* view of the mind that regards mental contents as nonmaterial and so causally impotent by-products of physical processes, to the extent that they exist at all. If no mental contents have causal efficacy in behavior, then *a fortiori* thoughts, beliefs, deliberation, reflection and reasoning can have none. Reason as a source of moral motivation is ruled out by fiat. But I argue in Volume II, Chapter V that this inference could not provide support for the causal efficacy of desire without further argument to demonstrate that occurrent desires are or can be interpreted as exclusively physical, whereas occurrent thoughts and beliefs are exclusively mental. And I call into question whether it is possible to demonstrate this.

A second, companion dogma is that of *mind-body materialism*, which claims that only third-personally observable physical matter exists. This is the metaphysical bedrock on which attempts to reduce desire to the exclusively physical rely. Again my aim is not to deny the existence or causal efficacy of observable physical matter, even though the concept of physical matter is starting to look increasingly primitive from the perspective of theoretical physics. It is rather to make explicit what most contemporary moral philosophy takes for granted, i.e. that first-personally observable mental states exist just as robustly and efficaciously. I undertake this task as well in Volume II's Chapter V, along with a third with which it is intertwined, namely *behaviorism*, the view that there are no inner states – at least none worth scientific notice; and that only those expressed in overt behavior are of interest. This dogma has a particularly robust pedigree, in psychology as well as in mid-century Positivism.

Conjointly these three dogmas bear a strong family resemblance to a fourth: what I describe in Volume II, Chapter V as the ideal of spontaneity and what neoclassical economics describes as the theory of *revealed preference*, i.e. that all inner states are revealed in overt physical behavior, whether verbal or nonverbal. In these cases as well, it is not the concentrated attention to physical behavior to which I object, but rather the doctrinal insistence that physical behavior is all there is. I argue that the anti-psychologistic constellation of epiphenomenalism, materialism and behaviorism was the expression of a reactive, mid-century aversion to psychological interiority consequent on the trauma of the second World War, which it is now time to re-evaluate.

A fifth dogma of Humeanism is the assumption that *sentential propositions are the fundamental units of meaning*, the intellective counterpart to the passional thesis targeted in this volume, that desire is the fundamental unit of motivation. If sentential propositions are the smallest units of meaning, then – unlike desire – the motivational efficacy of reason can reach no more deeply into the self than the motivational efficacy of sentences; and – unlike desire – can play no more atomistic, foundational or developmental role in the structure of the self. I argue against this assumption in Volume II, Chapter II. Obviously I do not reject the sentential proposition that sentential propositions have meaning. But I do reject the tacit assumption that they are the most basic units of meaning; and propose in Volume II, Chapters II and III a more fine-grained analysis of subsentential constituents to supplement it. Rethinking this fifth dogma is also necessary in order to show how reason can motivate action because if brain states are causally effective, and occurrent beliefs can be identified with brain states as mind-body materialism supposes, then occurrent beliefs can be causally effective after all. But not all occurrent beliefs can be formulated sententially because not all thoughts, ideas, images or associations are formulable in sentential terms. A more fine-grained analysis increases the chances of tracking observationally those intellective, non-passional mental events that occurrently precipitate action.

Connected with this is a sixth dogma of Humeanism, *the "is-ought" distinction* that confines truth to the former, descriptive realm and relegates to the latter, prescriptive realm the expression of emotions and attitudes. If meaning is located in sentential propositions, and meaningful sentential propositions can refer only to physical states of affairs, then in the end, no verbally expressed act of human intellection that does not refer to a physical state of affairs can be meaningful. I argue against this dogma in Volume II, Chapters V and IX, that so-called prescriptive sentences – i.e. commands and imperatives – are in fact categorical declarative sentences of the ordinary kind that are descriptive and explanatory of ideal states of affairs; and so are syntactically and epistemically on a par with those sentences descriptive and explanatory of non-ideal physical realities. Again the aim is to supplement rather than replace the canonical assumption.

These six dogmas interlock with those two models which conjoin to define explicitly *the Humean conception of the self*. Here the implication arrows point in both directions. The Humean conception accepts desire as the foundational element in both its model of motivation and its model of rationality. In order to extend its reach as an explanatory paradigm, it interprets desire behavioristically, as revealed preference theory requires. In so doing, it links expressions of desire exclusively with physical behavior, and so gives indirect support both to mind-body materialism and to epiphenomenalism. In rejecting the phenomenal and first-personal, it confines meaning to verbal behavior and the referents of language to physical states of

affairs, thus providing support both to the primacy of sentential propositions as the basic units of meaning, and to the is-ought distinction as confining meaning to the physically verifiable.

Conversely, these six dogmas lend support to the Humean conception. Behaviorism implies an interpretation of desire of the sort that revealed preference theory supplies. Epiphenomenalism and mind-body materialism together underwrite the interpretation of such desires as exclusively physical, and the interpretation of rational judgments about ideal and therefore non-verifiable states of affairs as both causally impotent and meaningless. This interpretation is given further credibility by the presupposition that sentential propositions that refer to physically verifiable states of affairs are the fundamental units of meaning.

If only third-personally observable physical behavior exists, then first-personal mental states do not. Rather, they are manifested in verbal and other physical behavior to the extent that they exist at all. If the mind is epiphenomenal and causally ineffectual, then verbal behavior that purports to express thoughts and beliefs in sentential propositions not only manifests epiphenomenal and causally ineffectual mental states of the agent, but also communicates them in an epiphenomenal and causally ineffectual manner. *A fortiori*, verbal behavior that purports to express rational thoughts and beliefs in sentential propositions that refer to ideal states of affairs manifests epiphenomenal and causally ineffectual states of the agent, communicates them epiphenomenally and ineffectually, and refers to nothing. Therefore reason is causally, i.e. motivationally ineffectual both first- and second-personally.

So not only are rational principles impotent to motivate our behavior first-personally. In addition, second-personal appeals to reason in others are impotent to motivate their behavior. Then in particular, the second-personal appeals to reason that form the foundation of philosophical practice in the Socratic metaethical tradition are in theory incapable of doing the job to which they purport to be committed. Hence philosophical practice itself as traditionally self-represented is without practical effect. Moreover, if all our actions seek to satisfy our desires, then maximizing the satisfaction of desire, i.e. utility, is our only final end, and this end is revealed in the physical behavior in which we engage. Then in particular, our physical behavior of, for example, analyzing, arguing, criticizing, theorizing and so on, maximizes our desires to do those things, and supplies some of the more innocuous reasons why we do philosophy. Similarly, physical behavior that maximizes the satisfaction of our desires to win, shine, show off, acquire power, or subjugate others supplies some of the more noxious ones. The transpersonally rational ideal of gaining reflective consensus on a transpersonally justifiable ethics, politics or society has nothing to do with it.

Taken together, then, these seven reductive dogmas – the Humean conception and the six metaphysical dogmas with which it is interdependent – exclude by fiat the very possibility that anything other than desire might be conceptually or psychologically significant for moral theory; indeed, that any conative state of the agent other than desire might truly be said to exist. It thus makes the case for egocentric rationality by denying not only the philosophical legitimacy but also the metaphysical existence of the cognitive and behavioral capacities that constitute transpersonal rationality. Under the weight of these radically exclusionary and repressive dogmas, it is little wonder that Kantians seem at a disadvantage in making their case.

These overly restrictive and reductive dogmas are examples of what I describe in Volume II, Chapter VII as *pseudorationality*. That is, they deny, dissociate or rationalize the exclusion of the very data of moral experience that are most in need of analysis and explanation, in order to preserve the illusion of rational intelligibility for those which remain. In Chapters VII, VIII, X, and XI respectively of Volume II, I offer four detailed test cases of pseudorationality, of increasing degrees of seriousness, realism and applicability to real-life circumstances, which illustrate the problems – for self-knowledge, knowledge of others and of the world, and for a realistic and effective moral response to them – that attachment to the mere illusion of rational intelligibility can precipitate. The price of attachment to these seven reductive dogmas of Humeanism is to leave unresolved the pressing problems of moral motivation and rational justification with which most Socratic metaethicists, not only those who are Humean devotees, are justifiably preoccupied. I have tried to suggest – and in Volume II try to demonstrate – that if we want to resolve them, we must leave all of these dogmas behind.

Bibliography

Aiken, Henry David, "An Interpretation of Hume's Theory of the Place of Reason in Ethics and Politics," *Ethics 90* (October 1979)

Allais, Maurice, "Fondements d'une Théorie Positive des Choix Comportant un Risque et Critique des Postulats et Axiomes de L'Ecole Americaine,"Memoir III of *Econometrie XL* (1953), 257-332 (Colloques Internationaux du Centre National de la Recherche Scientifique, Paris), translated as "Foundations of a Positive Theory of Choice Involving Risk and a Criticism of the Postulates and Axioms of the American School," in Maurice Allais and Ole Hagen, Eds. *Expected Utility and the Allais Paradox* (Dordrecht, Holland: D. Reidel, 1979), 27-146

_____ and Hagen, Ole, Eds. *Expected Utility and the Allais Paradox* (Dordrecht, Holland: D. Reidel, 1979)

Alexander, Peter, "Rational Behavior and Psychoanalytic Explanation," in N. S. Care and C. Landesman, Eds. *Readings in the Theory of Action* (Bloomington, Ind.: Indiana University Press, 1969)

Allison, Henry, *Kant's Theory of Freedom* (Cambridge: Cambridge University Press, 1990)

Anderson, Elizabeth, *Value in Ethics and Economics* (Cambridge, Mass.: Harvard University Press, 1993)

Anscombe, G. E. M., "Modern Moral Philosophy," *Philosophy 33* (1958), pp. 1-19

Aristotle, *Nicomachean Ethics*, trans. Terence Irwin (Indianapolis: Hackett, 1985)

Armstrong, D. M., *Belief, Truth and Knowledge* (London: Cambridge University Press, 1973)

Audi, Robert, "Psychoanalytic Explanation and the Concept of Rational Action," *The Monist 56* (1972), 444-464

Austin, J. L., "A Plea for Excuses," in *Philosophical Papers*, Ed. J. O. Urmson and G. J. Warnock (New York: Oxford University Press, 1970), 175-204

Baier, Annette, "Hume's Analysis of Pride," *The Journal of Philosophy LXXV*, 1 (January 1978), 27-40

_____, *A Progress of Sentiments: Reflections on Hume's Treatise* (Cambridge, Mass.: Harvard University Press, 1991)

_____, *Moral Prejudices* (Cambridge, Mass.: Harvard University Press, 1994)

_____, "Note on Justice, Care, and Immigration Policy," *Hypatia 10*, 2 (Spring 1995), 150-152

Baron, Marcia, "Hume's Calm Passions," (M. A. Thesis, The University of North Carolina at Chapel Hill, 1978)

_____, "The Alleged Repugnance of Acting from Duty," *The Journal of Philosophy LXXXI*, 4 (April 1984)

_____, *Kantian Ethics Almost without Apology* (Ithaca: Cornell University Press, 1995)

Bartlett, Donald L. and Steele, James B., *Empire: The Life, Legend and Madness of Howard Hughes* (New York: W. W. Norton and Company, 1979)

Benacerraf, Paul, "Mathematical Truth," *The Journal of Philosophy LXX*, l9, November 8, l973

Benn, S. I. and Gauss, G. F., "Practical Rationality and Commitment," *American Philosophical Quarterly 23*, 3 (July 1986), 255-266

Bennett, Jonathan, *Rationality* (London: Routledge and Kegan Paul Ltd., 1964)

_____, "Whatever the Consequences," *Analysis 26* (1966), pp. 83-102

Bentham, Jeremy, *Introduction to the Principles of Morals and Legislation*, Ed. J. H. Burns and H. L. A. Hart (London: Athlone, l970)

van Benthem, Johan and Liu, Fenrong, "Dynamic Logic of Preference Upgrade," *Journal of Applied Non-Classical Logics 14*, 2 (2004), 1 – 26

Bhagavadgita with the commentary of Sankaracarya, trans. Swami Gambhirananda (Calcutta: Advaita Ashrama, 1995)

The Bhagavad Gita with the Commentary of Sri Sankaracharya, trans. Alladhi Mahadeva Sastry (Madras: Samata Books, 1995)

The Bhagavad Gita, trans. Winthrop Sargeant (Albany: State University of New York Press, 1994)

Song of God: The Bhagavad Gita, trans. Swami Prabhavananda and Christopher Isherwood (New York: Mentor, 1972)

The Bhagavadgita, trans. S. Radhakrishnan (New York: Harper Torchbooks, 1973)

The Bhagavad-Gita: Krishna's Counsel in Time of War, trans. Barbara Stoler Miller ((New York: Bantam Books, 1986)

The Bhagavad Gita, trans. Juan Mascaro (London: Penguin Classics, 1962)

Bishop Butler, *Fifteen Sermons*, Sermon XI, 415; reprinted in *The British Moralists 1650-1800, Volume I: Hobbes-Gay*, Ed. D. D. Raphael (Oxford, The Clarendon Press, 1969)

Blackburn, Simon, *Ruling Passions* (Oxford: Oxford University Press, 2000)

_____, *Lust* (New York: Oxford University Press/ The New York Public Library, 2004)

Blum, Lawrence, *Friendship, Altruism and Morality* (Boston: Routledge and Kegan Paul, 1980)

Bolker, Ethan D., "A Simultaneous Axiomatization of Utility and Subjective Probability," *Philosophy of Science 34* (1967), 333-340

_____, "An Existence Theorem for the Logic of Decision," *Philosophy of Science 67* (2000), S14-S17

Brandom, Robert B., *Making It Explicit: Reasoning, Representing, and Discursive Commitment* (Cambridge, Mass.: Harvard University Press, 1994)

_____, *Articulating Reasons: An Introduction to Inferentialism* (Cambridge, Mass.: Harvard University Press, 2001)

Brandt, Richard B., *Ethical Theory* (Englewood Cliffs, N.J.: Prentice-Hall, 1959)

_____, "A Utilitarian Theory of Excuses," *The Philosophical Review LXXVII,* 3 (1969), pp. 337-61

_____, "Rational Desire," APA Western Division Presidential Address, *Proceedings and Addresses of the American Philosophical Association XLIII* (1969-1970), 43-64

_____, "Traits of Character: A Conceptual Analysis," *American Philosophical Quarterly 7*, 1 (January 1970)

_____, *A Theory of the Good and the Right* (Oxford: Clarendon Press, 1979)

_____ and Kim, Jaegwon, "Wants as Explanations of Action," *The Journal of Philosophy LX* (1963), 425-35; reprinted in N. S. Care and C. Landesman, Eds. *Readings in the Theory of Action* (Bloomington, Ind.: Indiana University Press, 1969), 199-213

Bratman, Michael, "Two Faces of Intention," *Philosophical Review XCIII,* 3 (July 1984), 375-405

_____, "Davidson's Theory of Intention," in Bruce Vermazen and Merrill B. Hintikka, Eds. *Essays on Davidson: Actions and Events* (Oxford: Clarendon Press, 1985), 13-26

Bromberg, Philip, "'Speak up that I may see you: Some reflections on dissociation, reality and psychoanalytic listening," *Psychoanalytic Dialogues 4* (1994), 517-547

Broome, John, "Utilitarianism and Expected Utility," *The Journal of Philosophy LXXXIV*, 8 (August 1987), 405-422

_____, "Rationality and the Sure-Thing Principle," in *Thoughtful Economic Man*, edited by Gay Meeks, Cambridge University Press, 1991, pp. 74–102

Care, N. S. and Landesman, C., Eds. *Readings in the Theory of Action* (Bloomington: Indiana University Press, 1969)

Cavalli-Sforza, L. L. and Bodmer, W. F., *The Genetics of Human Populations* (San Francisco: W. H. Freeman and Co., 1971)

Chisholm, Roderick, *Person and Object: A Metaphysical Study* (La Salle, Ill.: Open Court, 1976)

Cioffi, Frank, "Freud and the Idea of a Pseudo-Science," in Robert Borger and Frank Cioffi, *Explanation in the Behavioral Sciences* (Cambridge: Cambridge University Press, 1970)

Clarke, Samuel, *A Discourse Concerning the Unchangeable Obligations of Natural Religion*, Ed. L. A. Selby-Bigge, *The British Moralists, Vol. II* (New York: Dover, 1965)

Coase, Ronald, "The Problem of Social Cost," *Journal of Law and Economics 3* (1960), 1-44

_____, "Durability and Monopoly," *Journal of Law and Economics 15*, 1 (April 1972), 143-149

Cohen, L. Jonathan, "On the Psychology of Prediction: Whose is the Fallacy?" *Cognition 7* (1979), 385-407

_____, "Can Human Irrationality be Experimentally Demonstrated?" *Behavioral and Brain Sciences 4* (1981), 317-370

Custance, John, "The Universe of Bliss and the Universe of Horror: A Description of a Manic-Depressive Psychosis," in Bert Kaplan, Ed. *The Inner World of Mental Illness* (New York: Harper and Row, 1964)

Cyert, Richard M. and DeGroot, Morris H., "Adaptive Utilities," in Allais, Maurice and Hagen, Ole, Eds. *Expected Utility and the Allais Paradox* (Dordrecht, Holland: D. Reidel, 1979)

DSM III: Diagnostic and Statistical Manual of Mental Disorders, Third Edition (Washington, D.C.: The American Psychiatric Association, 1980)

Daniels, Norman, Ed. Reading *Rawls* (New York: Basic Books, Inc., 1974)

Danto, Arthur, "Basic Actions," in Care, N. S. and Landesman, C., Eds. *Readings in the Theory of Action* (Bloomington: Indiana University Press, 1969)

Darwall, Stephen, *Impartial Reason* (Ithaca, New York: Cornell University Press, 1983)

Davidson, Donald, "On the Very Idea of a Conceptual Scheme," APA Presidential Address, *Proceedings and Addresses of the American Philosophical Association 47* (1974)

_____, "Psychology as Philosophy," in *Essays on Actions and Events* (Oxford: Clarendon Press, 1980)

_____, "How is Weakness of the Will Possible?" in _____

_____, McKinsey, J. C. C., and Suppes, Patrick, "Outlines of a Formal Theory of Value, I," *Philosophy of Science 22,* 2 (April 1955), 140-160

_____, Siegel, Sidney, and Suppes, Patrick, "Some Experiments and Related Theory on the Measurement of Utility and Subjective Probability," Applied Mathematics and Statistics Laboratory, *Technical Report 1*, Stanford University, Stanford, Cal., August 15, 1955

Davis, F. James, *Who Is Black?* (University Park: Pennsylvania State University Press, 1991)

Davis, Wayne, "A Theory of Happiness," *American Philosophical Quarterly 18*, 2 (April 1981), 111-119

_____, "Pleasure and Happiness," *Philosophical Studies 39* (1981), 305-317

Dennett, Daniel, "Intentional Systems," *The Journal of Philosophy LXIII*, 4 (February 25, 1971), 87-106

Dent, N. J. H., *The Moral Psychology of the Virtues* (Cambridge: Cambridge University Press, 1984)

Dominguez, Virginia R., *White By Definition: Social Classification in Creole Louisiana* (New Brunswick: Rutgers University Press, 1986)

Douglas, Mary, *Purity and Danger* (London: Routledge and Kegan Paul, 1966)

Dummett, Michael, *Frege's Philosophy of Language* (New York: Harper and Row, 1973)

Dworkin, Ronald, "The Original Position" (*University of Chicago Law Review 40*, 3 (Spring 1973), 500-533

_____, *Taking Rights Seriously* (Cambridge, Mass.: Harvard University Press, 1977)

Edelman, Gerald M., *Neural Darwinism: The Theory of Neuronal Group Selection* (New York: Basic Books, 1987)

_____, *The Remembered Present: A Biological Theory of Consciousness* (New York: Basic Books, 1989)

Edgeworth, Francis Ysidro, *Mathematical Psychics and Other Essays* (San Diego: James and Gordon, 1995)

Edwards, Ward, "Probability- Preferences in Gambling," *American Journal of Psychology 66,* 3 (1953), 349-364

_____, "Experiments on Economic Decision-Making in Gambling Situations," *Econometrica 21* (1953), 349-350

_____, "Probability Preferences Among Bets with Differing Expected Values," *American Journal of Psychology 67* (1954), 56-67

_____, "The Reliability of Probability Preferences," *American Journal of Psychology 67* (1954), 68-95

_____, "The Theory of Decision-Making," *Psychological Bulletin 51*, 4 (1954), 380-417

Elster, Jon, *Ulysses and the Sirens: Studies in Rationality and Irrationality* (New York: Cambridge University Press, 1979)

Epictetus, *Enchiridion* LI. trans. P.E. Matheson (Oxford: Clarendon Press), reprinted in Jason L. Saunders, Ed. *Greek and Roman Philosophy after Aristotle* (New York: The Free Press, 1966); and trans. George Long (Chicago: Henry Regnery Co., 1956)

Erwin, Edward, "The Truth about Psychoanalysis," *The Journal of Philosophy LXXXVIII*, 10 (October 1981), 549-560

Evans-Pritchard, E. E., *The Nuer: A Description of the Modes of Livelihood and Political Institutions of a Nilotic People* (Oxford: Clarendon Press, 1940)

Falk, W. D., "'Ought' and Motivation," *Proceedings of the Aristotelian Society*, New Series, NLVIII (1947-1948), 111-138

_____, "Morality, Self, and Others," in Judith J. Thomson and Gerald Dworkin, Eds., *Ethics* (New York: Harper and Row, 1968); reprinted in Hector-Neri Castaneda and George Nakhnikian, Eds. *Morality and the Language of Conduct* (Detroit: Wayne State University press, 1963)

_____, "Hume on Practical Reason," *Philosophical Studies 27* (1975), 1-18

Farnsworth, Clyde H., "Survey of Whistle Blowers Finds Retaliation but Few Regrets," *The New York Times* (Sunday, February 22, 1987), 22

_____, "In Defense of the Government's Whistle Blowers," *The New York Times* (Tuesday, July 26, 1988), page B6

Farrell, B. A., "The Criteria for a Psychoanalytic Explanation," *Proceedings of the Aristotelian Society, Supplementary Volume XXXVI* (1962); reprinted in D. Gustafson, Ed. *Philosophical Psychology* (New York: Doubleday, Inc., 1964)

Fechner, G. T., *Elements of Psychophysics*, Vol. I, Trans. H. E. Adler, Ed. E. G. Boring and D. Howes (New York: Holt, Rinehart and Winston, 1966)

Feinberg, Joel, "Action and Responsibility," in *Doing and Deserving* (Princeton, N. J.: Princeton University Press, 1970)

_____, "The Idea of a Free Man," in *Rights, Justice, and the Bounds of Liberty* (Princeton: Princeton University Press, 1980)

_____, "Psychological Egoism," in Joel Feinberg and Russ Shafer-Landau, Eds., *Reason and Responsibility: Readings in Some Basic Problems of Philosophy* (Belmont, Cal.: Wadsworth Publishing Company, 1998), 493-505

Fishburn, Peter C., "On the Nature of Expected Utility," in Allais, Maurice and Hagen, Ole, Eds. *Expected Utility and the Allais Paradox* (Dordrecht, Holland: D. Reidel, 1979)

Fodor Jerry A., "Language, Thought and Compositionality," *Mind & Language 16*, 1 (February 2001), 1-15

_____ and Lepore, Ernest, *The Compositionality Papers* (New York: Oxford University Press, 2002)

Foot, Phillippa, "Morality as a System of Hypothetic Imperatives," *The Philosophical Review LXXXI* (1972), 306-16

Frankena, William, *Ethics*, Second Edition (Englewood Cliffs, N.J.: Prentice-Hall, 1973)

Frankfurt, Harry, "Freedom of the Will and the Concept of a Person," *The Journal of Philosophy LXVIII*, 1 (January 1971), 5-20

_____, "Identification and Externality," in Amelie O. Rorty, Ed. *The Identities of Persons* (Berkeley: University of California Press, 1976)

_____, "Rationality and the Unthinkable," in *The Importance of What We Care About: Philosophical Essays* (New York: Cambridge University Press, 1989), 177-190

Gauthier, David, "Justice and Natural Endowment: Toward a Critique of Rawls' Ideological Framework," *Social Theory and Practice 3* (1975), 3-26

_____, "The Social Contract as Ideology," *Philosophy and Public Affairs 6* (1977), 130-164

_____, "Economic Rationality and Moral Side-Constraints," *Midwest Studies in Philosophy III: Studies in Ethical Theory* (Minneapolis: University of Minnesota Press, 1978)

_____, "The Incompleat Egoist: From Rational Choice to Moral Theory," *The Tanner Lectures* (Palo Alto: Stanford University Press, 1983)

_____, *Morals by Agreement* (New York: Oxford University Press, 1985)

Gautier, Theophile, "The Nights of Cleopatra," in *Mademoiselle de Maupin* (New York: Modern Library, 1949)

Gewirth, Alan, *Reason and Morality* (Chicago: University of Chicago Press, 1978)

Gibbard, Allan, "Utilitarianisms and Coordinations" (Ph.D. diss., Harvard University, 1971)

_____, "Interpersonal Comparisons: Preference, Good, and the Intrinsic Reward of a Life," in *Foundations of Social Choice Theory,* Edited by Jon Elster and Aanund Hylland (New York: Cambridge University Press, 1989)

_____, *Wise Choices, Apt Feelings: A Theory of Normative Judgment* (Cambridge, Mass.: Harvard University Press, 1990)

Gigerenzer, Gerd, "Fast and Frugal Heuristics: The Tools of Bounded Rationality," in D. Koehler and N. Harvey, Eds. *Blackwell Handbook of Judgment and Decision-Making* (Oxford, UK: Blackwell, 2004), 62-88.

_____, "Bounded and Rational," in R. J. Stainton, Ed. *Contemporary Debates in Cognitive Science* (Oxford, UK: Blackwell, 2006), 115-133.

Gilligan, Carol, *In a Different Voice: Psychological Theory and Women's Development* (Cambridge, Mass.: Harvard University Press, 1982)

Glazer, Myron Peretz and Glazer, Penina Migdal, *The Whistleblowers: Exposing Corruption in Government and Industry* (New York: Basic Books, 1989)

Goldman, Alvin, *A Theory of Human Action* (New Jersey: Prentice-Hall, 1970)

Gorovitz, Samuel, "The Saint Petersburg Puzzle" in Allais, Maurice and Hagen, Ole, Eds. *Expected Utility and the Allais Paradox* (Dordrecht, Holland: D. Reidel, 1979)

Grandy, Richard, Ed. *Theories and Observation in Science* (Englewood, N.J.: Prentice-Hall, 1973)

Grice, H. P., "Meaning," *Philosophical Review 66* (1957): 377-88

Grünbaum, Adolph, "How Scientific is Psychoanalysis?" in Raphael Stern, Louise S. Horowitz, and Jack Lynes, Eds., *Science and Psychotherapy* (New York: Haven, 1977)

_____, "Is Freudian Psychoanalytic Theory Pseudo-Scientific by Karl Popper's Criterion of Demarcation?" *Americal Philosophical Quarterly XVI,* 2 (April 1979), 131-141

_____, "Epistemological Liabilities of the Clinical Appraisal of Psychoanalytic Theory," *Nous XIV*, 3 (September 1980), 307-385

Habermas, Jürgen, "Reconciliation Through the Public Use of Reason: Remarks on John Rawls's Political Liberalism," *The Journal of Philosophy XCII*, 3 (March 1995), 109-131

_____, *The Inclusion of the Other: Studies in Political Theory*, trans. Ciaran Cronin (Cambridge, Mass.: MIT Press, 1998)

_____, *Moral Consciousness and Communicative Action*, trans. Christian Lenhardt and Shierry Weber Nicholsen (Cambridge, Mass., MIT Press, 1999)

Hammond, Peter, "Changing Tastes and Coherent Dynamic Choice," *The Review of Economic Studies 43* (1976) , 159-73

_____, "Dynamic Restrictions on Metastatic Choice," *Economica 44* (1977), 337-50

_____, "Consequential Foundations for Expected Utility," *Theory and Decision 25* (1988), 25-78

Hampshire, Stuart, "Liberator, Up to a Point," *The New York Review of Books XXXIV*, 5 (March 26, 1987)

Hanna, Robert, "Rationality and the Ethics of Logic," *The Journal of Philosophy CIII*, 2 (February 2006), 67-100.

Hanson, Norwood, "Observation," in Richard Grandy, Ed. *Theories and Observation in Science* (Englewood, N.J.: Prentice-Hall, 1973), 129-146

Hardie, W. F. R. "The Final Good in Aristotle's Ethics," *Philosophy XL* (l965), 277-295

Harman, Gilbert, "Moral Relativism Defended," *The Philosophical Review LXXXIV* (1975), 3-22

Harsanyi, John C., "Advances in Understanding Rational Behavior," in John Harsanyi, *Essays on Ethics, Social Behavior, and Scientific Explanation* (Dordrecht: D. Reidel, 1976)

_____, *Rational Behavior and Bargaining Equilibrium in Games and Social Situations* (New York: Cambridge University Press, 1977)

_____, "Morality and the Theory of Rational Behavior," *Social Research 44* (1977), 623-656

Hegel, G. W. F. *The Philosophy of Right,* trans. T. M. Knox (New York: 1975)

Henrich, Dieter, *The Unity of Reason: Essays on Kant's Philosophy*, Ed. Richard Velkey (Cambridge, Mass.: Harvard University Press, 1994)

Herman, Barbara, "On the Value of Acting from the Motive of Duty," *Philosophical Review 66* (1981): 359-382

_____, *The Practice of Moral Judgment* (Cambridge, Mass.: Harvard University Press, 1993)

Hilts, Philip J., "Why Whistle-Blowers Can Seem a Little Crazy," *The New York Times* (Sunday, June 13, 1993), Section 4, page 6

_____, *Smokescreen: The Truth Behind the Tobacco Industry Cover-up* (New York: Addison-Wesley Publishing Company, Inc., 1996)

Hirschman, Albert, *The Passions and the Interests* (Princeton: Princeton University Press, 1977)

Hobbes, Thomas, *Leviathan*, Ed. Michael Oakeshott (New York: Collier, 1977)

Hodgson, D.H., *Consequences of Utilitarianism* (Oxford: Clarendon Press, 1967)

Horton, Robin, "African Traditional Thought and Western Science," in Bryan Wilson, Ed., *Rationality* (New York: Harper and Row, 1970), 131-171

Howell, Robert, *Kant's Transcendental Deduction: An Analysis of Main Themes in His Critical Philosophy* (Dordrecht: Kluwer Academic Publishers, 1992)

Hoyningen-Huene, Paul, "Systematizität: Die Natur der Wissenschaft," unpublished manuscript (delivered to die Gesellschaft für analytische Philosophie 6th Congress, Freie Universität Berlin, September 2006)

Humberstone, I. L., ("Wanting as Believing," *The Canadian Journal of Philosophy 17*, 1 (March 1987), 49-62

Hume, David, *A Treatise of Human Nature*, Ed. L. A. Selby-Bigge (Oxford: Clarendon Press, 1968)

_____, *Enquiry Concerning the Principles of Morals*, Ed. J. Schneewind (Indianapolis, Ind.: Hackett Publishing Co.)

_____, *Enquiry Concerning the Human Understanding and Concerning the Principles of Morals*, Ed. L. A. Selby-Bigge, Second Edition (Oxford: Clarendon Press, 1966)

_____, *Essays: Moral Political and Literary*, Ed. Eugene F. Miller (Indianapolis, Ind.: Liberty Classics, 1985)

Hunt, Liz, "Whistleblowers 'put their health under threat'," *The Independent* (Friday, 10 September 1993), Section 1, p. 6

Hutcheson, Francis, *Illustrations on the Moral Sense*, Ed. Bernard Peach (Cambridge, Mass.: Belknap Press of Harvard University, 1971)

_____, "An Inquiry Concerning Moral Good and Evil," in Raphael, D. D., Ed., *The British Moralists 1650-1800, Volume I: Hobbes-Gay*. (Oxford: The Clarendon Press, 1969)

Jacobs, Jane, *Systems of Survival: A Dialogue on the Moral Foundations of Commerce and Politics* (New York: Random House, 1992)

Jeffrey, Richard C., *The Logic of Decision*, Second Edition (Chicago: University of Chicago Press, 1983)

de Jongh, Dick and Liu, Fenrong, "Optimality, Belief and Preference," in *Proceedings of the Workshop on Rationality and Knowledge, ESSLLI 2006*, Ed. Sergei Artemov and Rohit Parikh (Amsterdam: University of Amsterdam, 2006), 1 – 12. Delivered to the *Models of Preference Change* Workshop, Freie Universität Berlin, 15 September 2006

Kahn, Virginia Munger, "Brokers Making Amends for Trading Problems," *The New York Times* (Sunday, November 2, 1997), Money and Business Section, 8

Kant, Immanuel, *Kritik der Reinen Vernunft*, *Kant's gesammelte Schriften*, herausg. königlich Preußischen bzw. Deutschen Akademie der Wissenschaften (Berlin, 1911; New York: Walter de Gruyter), Vols. 3 [B Edition] and 4 [A Edition]

_____, *Kritik der Reinen Vernunft*, Herausg. Raymund Schmidt (Hamburg: Felix Meiner Verlag, 1976)

_____*The Critique of Pure Reason*, trans. Paul Guyer and Allen W. Wood (New York, N.Y.: Cambridge University Press, 1998)

_____, *The Critique of Pure Reason*, trans. Norman Kemp Smith (New York, N.Y.: St. Martin's Press, 1970)

_____, *Prolegomena zu einer jeden künftigen Metaphysik, die als Wissenschft wird auftreten können*, *Kant's gesammelte Schriften*, herausg. königlich Preußischen bzw. Deutschen Akademie der Wissenschaften (Berlin, 1911; New York: Walter de Gruyter), Vol. 4

_____, *Prolegomena to Any Future Metaphysics,* trans. Lewis White Beck (New York, N.Y.: Bobbs-Merrill, 1950)

_____, *Grundlegung zur Metaphysik der Sitten*, *Kant's gesammelte Schriften*, herausg. königlich Preußischen bzw. Deutschen Akademie der Wissenschaften (Berlin, 1911; New York: Walter de Gruyter), Vol. 4

_____, *Grundlegung zur Metaphysik der Sitten*, Herausg. von Karl Vorländer (Hamburg: Felix Meiner Verlag, (1965)

_____, *Fundamental Principles of the Metaphysic of Morals*, trans. Thomas K. Abbott (New York: Bobbs-Merrill, 1949)

_____, *Foundations of the Metaphysics of Morals*, trans. Lewis White Beck; text and critical essays edited by Robert Paul Wolff (New York: Bobbs-Merrill, 1969)

_____, *Grounding for the Metaphysics of Morals*, trans. James W. Ellington (Indianapolis: Hackett 1981)

_____, *Groundwork of the Metaphysic of Morals*, trans. H. J. Paton (New York, N.Y.: Harper Torchbooks, 1964)

_____, *Groundwork for the Metaphysics of Morals*, ed. and trans. Allen W. Wood (New Haven: Yale University Press, 2002) with critical essays

_____, *Kritik der praktischen Vernunft*, *Kant's gesammelte Schriften*, herausg. königlich Preußischen bzw. Deutschen Akademie der Wissenschaften (Berlin, 1911; New York: Walter de Gruyter), Vol. 5

_____, *Kritik der praktischen Vernunft*, Herausg. Karl Vorländer (Hamburg: Felix Meiner Verlag, 1974)

_____, *The Critique of Practical Reason,* trans. Lewis White Beck (New York, N.Y.: Bobbs-Merrill, 1956)

_____, *Critique of Practical Reason*, trans. Mary J. Gregor (New York: Cambridge University Press, 1997)

_____, *Die Religion innerhalb der Grenzen der bloßen Vernunft*, *Kant's gesammelte Schriften*, herausg. königlich Preußischen bzw. Deutschen Akademie der Wissenschaften (Berlin, 1911; New York: Walter de Gruyter), Vol. 6

_____, *Die Religion innerhalb der Grenzen der bloßen Vernunft*, Herausg. Karl Vorländer (Hamburg: Felix Meiner Verlag, 1978)

_____, *Religion Within the Limits of Reason Alone*, trans. T. M. Greene and H. H. Hudson (New York, N.Y.: Harper Torchbooks, 1960)

_____, *Metaphysik der Sitten, Kant's gesammelte Schriften*, herausg. königlich Preußischen bzw. Deutschen Akademie der Wissenschaften (Berlin, 1911; New York: Walter de Gruyter), Vol. 6

_____, *Metaphysik der Sitten,* Herausg. Karl Vorländer (Hamburg: Felix Meiner Verlag, 1966)

_____, *The Metaphysical Elements of Justice: Part I of the Metaphysics of Morals*, trans. John Ladd (New York: Bobbs-Merrill, 1965)

_____, *The Doctrine of Virtue: Part II of The Metaphysic of Morals*, trans. Mary J. Gregor (Philadelphia: University of Pennsylvania Press, 1971)

_____, *The Metaphysics of Morals*, trans. Mary J. Gregor (New York, N.Y.: Cambridge University Press, 1991)

_____, *Logik*, *Kant's gesammelte Schriften*, herausg. königlich Preußischen bzw. Deutschen Akademie der Wissenschaften (Berlin, 1911; New York: Walter de Gruyter),Vol. 9

_____, *Logic*, trans. Robert Hartman and Wolfgang Schwarz (New York: Bobbs-Merrill, 1974)

_____, *Kritik der Urteilskraft*, *Kant's gesammelte Schriften*, herausg. königlich Preußischen bzw. Deutschen Akademie der Wissenschaften (Berlin, 1911; New York: Walter de Gruyter), Volume 5

_____, *Kritik der Urteilskraft*, herausg. von Karl Vorländer (Hamburg: Felix Meiner Verlag, 1974)

_____, *Critique of Judgment*, trans. J. H. Bernard (New York: Hafner Publishing Company, 1972)

_____, *The Critique of Judgement*, trans. James Creed Meredith (Oxford: Oxford University Press, 1973)

_____, *Critique of Judgment*, trans. Werner S. Pluhar (Indianapolis: Hackett, 1987)

_____, *Erste Einleitung in die Kritik der Urteilskraft*, herausg. von Gerhard Lehmann (Hamburg: Felix Meiner Verlag, 1977)

_____, *First Introduction to the Critique of Judgment*, trans. James Haden (Indianapolis: Bobb-Merrill, 1965)

Kaplan, Bert, Ed. *The Inner World of Mental Illness* (New York: Harper and Row, 1964)

Kaplan, Mark, *Decision Theory as Philosophy* (New York: Cambridge, 1996)

_____, "Decision Theory and Epistemology," Section III, in Paul K. Moser, Ed., *The Oxford Handbook of Epistemology* (New York: Oxford University Press, 2002)

Katona, George, "Rational Behavior and Economic Behavior," *Psychological Review 60*, 5 (1953), 307-318

Kernberg, Otto, *Borderline Conditions and Pathological Narcissism* (New York: J. Aronson, 1975)

_____, *Severe Personality Disorders* (New Haven: Yale University Press, 1984)

Keynes, John Maynard, *The Economic Consequences of the Peace* (Mineola, New York: Dover Publications, 2004; orig. London: Macmillan and Co., 1920)

_____, "My Early Beliefs," in *Two Memoirs* (New York: Augustus M. Kelley, 1949), 85 and 88

Kim, Jaegwon, "Noncausal Connections," *Nous 8* (1974), pp. 41-52

Kitcher, Patricia, *Kant's Transcendental Psychology* (New York: Oxford University Press, 1990)

Kleinfeld, N. R., "The Whistle Blowers' Morning After," *The New York Times* (Sunday, November 9, 1986), Section 3, page 1

Kluger, Richard, *Ashes to Ashes: America's Hundred-Year Cigarette War, the Public Health, and the Unabashed Triumph of Philip Morris* (New York: Alfred A. Knopf, 1996)

Kohlberg, Lawrence, "The Claim to Adequacy of a Highest Stage of Moral Judgment," *The Journal of Philosophy LXX*, 18 (October 25, 1973), 630-646

Koopmans, Tjalling, "Allocation of Resources and the Price System," in *Three Essays on the State of Economic Science* (New York: McGraw-Hill, 1957)

Kronman, Anthony, Unpublished comments on John Rawls, "The Basic Liberties and Their Priority," *The Tanner Lecture on Human Values,* delivered at the University of Michigan, April 1981.

Kubler, George, *The Shape of Time: Remarks on the History of Things* (New Haven and London: Yale University Press, 1962)

Kuhn, Thomas, *The Structure of Scientific Revolutions* (Chicago: The University of Chicago Press, 1970)

Kydd, Rachel, *Reason and Conduct in Hume's Treatise* (New York: Russell and Russell, 1964)

Leonard, William E., "Excerpt from *The Locomotive God,*" in Bert Kaplan, Ed. *The Inner World of Mental Illness* (New York: Harper and Row, 1964)

Levi, Isaac, *Hard Choices* (New York: Cambridge University Press, 1986)

Lewis, David, *Convention: A Philosophical Study* (Cambridge, Mass: Harvard University Press, 1969)

_____, "Utilitarianism and Truthfulness," *Australasian Journal of Philosophy 50*, 1 (1972), 17-19

_____,"Radical Interpretation," *Synthese 23* (1974), 331-44. Reprinted in *Philosophical Papers, Volume I* (New York: Oxford University Press, 1983), 108-121

_____, "Attitudes *De Dicto* and *De Se*," in *Philosophical Papers, Volume I* (New York: Oxford University Press, 1983)

_____, "Desire as Belief," *Mind 97*, 387 (July 1988), 323-332

Liebenstein, Harvey, *Beyond Economic Man* (Cambridge, Mass.: Harvard University Press, 1976)

Linder, Staffan B., *The Harried Leisure Class* (New York: Columbia University Press, 1970)

Little, I. M. D., "A Reformulation of the Theory of Consumer's Behavior," *Oxford Economic Papers I* (1949), 90-99

_____, *Critique of Welfare Economics* (New York: Oxford University Press, 1970)

Loar, Brian, "The Semantics of Singular Terms," *Philosophical Studies 30* (1976), 353-77

Longuenesse, Béatrice, *Kant and the Capacity to Judge: Sensibility and Discursivity in the Transcendental Analytic of the Critique of Pure Reason*, trans. Charles T. Wolfe (Princeton: Princeton University Press, 1998)

Luce, R. D. and Raiffa, Howard, *Games and Decisions* (New York: John Wiley and Sons, Inc., 1957)

Lyons, David, *Forms and Limits of Utilitarianism* (Oxford: Clarendon Press, 1965)

March, James G., "Bounded Rationality, Ambiguity, and the Engineering of Choice," *Bell Journal of Economics 9* (1978), 587-608

Marschak, J., "Utilities, Values, and Decision Makers," in Allais, Maurice and Hagen, Ole, Eds. *Expected Utility and the Allais Paradox* (Dordrecht, Holland: D. Reidel, 1979)

McClennen, Edward, *Rationality and Dynamic Choice: Foundational Explorations* (New York: Cambridge University Press, 1990)

_____, "Pragmatic Rationality and Rules," *Philosophy and Public Affairs 26*, 3 (Summer 1997), 210-258

McCloskey, H. M., "A Note on Utilitarian Punishment," *Mind 72* (1963), 599

Mead, George Herbert, "Fragments on Ethics," in *Mind, Self and Society* (Chicago: University of Chicago Press, 1934), 379 ff.

Melden, Abraham Irving, *Free Action* (London: Routledge & Kegan Paul, 1961)

Meyer, Eugene and Covi, Lino,"The Experience of Depersonalization: A Written Report by a Patient," in Bert Kaplan, Ed. *The Inner World of Mental Illness* (New York: Harper and Row, 1964)

Milgram, Stanley, "Behavior Study of Obedience," *Journal of Abnormal and Social Psychology 67* (1963), 371 – 378

_____, *Obedience to Authority: An Experimental View* (New York: Harper/Collins, 1983)

Mill, John Stuart, *Utilitarianism*, Ed. George Sher (Cambridge: Hackett Publishing Co., 1979)

_____, *Utilitarianism* (New York, N.Y.: Bobbs-Merrill, 1979)

Miller, David, *Philosophy and Ideology in Hume's Political Thought* (Oxford: Clarendon Press, 1981)

Miller, Richard, "Rawls and Marxism," in Daniels, Norman, Ed. Reading *Rawls*, (New York: Basic Books, Inc., 1974)

_____, "Ways of Moral Learning," *The Philosophical Review XCIV*, 4 (October 1985), 507-556

Millgram, Elijah, "Does the Categorical Imperative Give Rise to a Contradiction in the Will?" *The Philosophical Review 112*, 4 (October 2003), 525 – 560

Mischel, Theodore, "Concerning Rational Behavior and Psychoanalytic Explanation," *Mind 74* (1965), 71-78

Moody, E., *The Logic of William of Ockham* (New York: Russell and Russell, 1965), 70-75)

Moore, G. E., *Principia Ethica* (Cambridge: Cambridge University Press, 1968)

Morgenstern, Oskar, "Thirteen Critical Points in Contemporary Economic Theory: An Interpretation," *Journal of Economic Literature 10* (1972), 1163-1189

_____, "Some Reflections on Utility," in Allais, Maurice and Hagen, Ole, Eds. *Expected Utility and the Allais Paradox* (Dordrecht, Holland: D. Reidel, 1979)

Mullane, Harvey, "Psychoanalytic Explanation and Rationality," *The Journal of Philosophy LXVIII*, 14 (1971), 413-426

Nagel, Thomas, *The Possibility of Altruism* (Oxford: Clarendon Press, 1970)

_____, "Rawls on Justice," *The Philosophical Review 87*, 2 (April 1973), 220-34; reprinted in *Reading Rawls*, Ed. Norman Daniels (New York: Basic Books, Inc., 1974)

_____, "Subjective and Objective," in *Mortal Questions* (Cambridge: Cambridge University Press, 1979)

_____, *The View From Nowhere* (New York: Oxford University Press, 1986)

Neely, Wright, "Freedom and Desire," *The Philosophical Review LXXXXIII,* 1 (January 1974), 32-54

Neisser, Ulric, "Cultural and Cognitive Discontinuity," in T. E. Gladwin and W. Sturtevant, Eds., *Anthropology and Human Behavior* (Washington, D. C.: Anthropological Society of Washington, 1962)

Nell [née O'Neill], Onora, *Acting on Principle: An Essay in Kantian Ethics* (New York: Columbia University Press, 1975)

Nietzsche, Friedrich, *On the Genealogy of Morals* and *Ecce Homo*, Trans. Walter Kaufmann and R. J. Hollingdale (New York: Vintage, 1967)

Nisbett, Richard E. and Wilson, Timothy, "Telling More than We Can Know: Verbal Reports on Mental Processes," *Psychological Review LXXXIV* (1977), 231-259

Nisbett, Richard E. and Ross, Lee, *Human Inference: Strategies and Shortcomings of Social Judgment* (Englewood Cliffs, N. J. Prentice-Hall, 1980)

Norton, David Fate, *David Hume: Common-Sense Moralist, Sceptical Metaphysician* (Princeton: Princeton University Press, 1982)

Nozick, Robert , *Anarchy, State and Utopia* (New York: Basic Books, 1974)

O'Neill, Onora, *Constructions of Reason* (Cambridge: Cambridge University Press, 1989)

_____, "Kant's Justice and Kantian Justice," in *The Bounds of Justice* (Cambridge: Cambridge University Press, 2000)

_____, "Autonomy: The Emperor's New Clothes," The Inaugural Address, *Proceedings of the Aristotelian Society, Supp. Vol. LXXVII* (2003), 1-21

_____, "Kantian Ethics," *Routledge Encyclopedia of Philosophy* (London: Routledge, 2000), 433

Paul, Jeffrey, Ed. *Reading Nozick* (Totowa, NJ: Rowman and Allenheld, 1981)

Peacocke, Christopher, "Intention and Akrasia," in Bruce Vermazen and Merrill B. Hintikka, Eds. *Essays on Davidson: Actions and Events* (Oxford: Clarendon Press, 1985) 51-74

Pelczynski, Z. A., Ed. *Hegel's Political Philosophy* (New York: Cambridge University Press, 1972)

Perry, John, Ed. *Personal Identity* (Los Angeles: University of California, 1975)

_____, "The Problem of the Essential Indexical," *Nous 13* (1979), 3-21

Pettit, Philip and Smith, Michael, "Backgrounding Desire," *The Philosophical Review XCIX*, 4 (October 1990), 565-592

Pieyre de Mandiargue, Andre, *The Margin*, Trans. Richard Howard (London: Calder and Boyars Ltd., 1969)

Piper, Adrian M. S., "Utility, Publicity and Manipulation," *Ethics 88*, 3 (April 1978), 189-206

_____,"Property and the Limits of the Self," *Political Theory 8*, 1 (February 1980), 39-64

_____, "A Distinction Without a Difference," *Midwest Studies in Philosophy VII: Social and Political Philosophy* (1982), 403-435

_____,"Two Conceptions of the Self," *Philosophical Studies 48*, 2 (September l985), 173-197; reprinted in *The Philosopher's Annual VIII* (1985), 222-246

_____, "Michael Slote's *Goods and Virtues,*" reviewed for *The Journal of Philosophy LXXXIII*, 8 (August 1986), 468-73

_____, "Instrumentalism, Objectivity, and Moral Justification," *American Philosophical Quarterly 23,* 4 (October 1986), 373-381

_____, "Moral Theory and Moral Alienation," *The Journal of Philosophy LXXXIV*, 2 (February 1987), 102-118

_____, "Personal Continuity and Instrumental Rationality in Rawls' Theory of Justice," *Social Theory and Practice 13*, 1 (Spring 1987), 49-76

_____, "Pseudorationality," in Amelie O. Rorty and Brian McLaughlin, Eds. *Perspectives on Self-Deception* (Los Angeles: University of California, 1988)

_____, "Hume on Rational Final Ends," *Philosophy Research Archives XIV* (1988-89), 193-228

_____, "'Seeing Things'," *Southern Journal of Philosophy XXIX, Supplementary Volume: Moral Epistemology* (1990), 29-60

_____, "Impartiality, Compassion, and Modal Imagination," *Ethics 101* (July 1991), 726 – 757

_____, "Xenophobia and Kantian Rationalism," *Philosophical Forum XXIV*, 1-3 (Fall-Spring 1992-93), 188-232. Reprinted in *Feminist Interpretations of Immanuel Kant*, Ed. Robin May Schott (University Park: Pennsylvania State University Press, 1997), 21-73; and in *African-American*

Perspectives and Philosophical Traditions, Ed. John P. Pittman (New York: Routledge, 1997)

_____, "Two Kinds of Discrimination," *Yale Journal of Criticism 6*, 1 (1993), 25-74. Reprinted in *Race and Racism*, ed. Bernard Boxill (Oxford: Oxford University Press), pp. 193-237

_____, "Making Sense of Value," *Ethics 106*, 2 (April 1996), 525-537

_____, "Kant on the Objectivity of the Moral Law," in Andrews Reath, Barbara Herman and Christine M. Korsgaard, Eds., *Reclaiming the History of Ethics: Essays for John Rawls* (New York: Cambridge University Press, 1997), 240-269

_____, "The Enterprise of Socratic Metaethics," in Naomi Zack, Ed., *Women of Color and Philosophy* (New York: Blackwell, 2000)

_____, "Kants intelligibler Standpunkt zum Handeln," in *Systematische Ethik mit Kant*, Eds. Hans-Ulrich Baumgarten and Carsten Held (München/Freiburg: 2001)

_____, "Letter to a Young Artist," *Art on Paper 9*, 6 (July/August 2005), 36-37; reprinted in Peter Nesbett and Sarah Andress, Eds. *Letters to a Young Artist,* (New York: Darte Publishing, 2006), 83-88

Plato, *Apology*, in *Euthyphro, Apology, Crito*, Trans. F. J. Church and Robert D. Cumming (New York: Bobbs-Merrill, 1956)

Platts, Mark, "Moral Reality and the End of Desire," in *Reference, Truth and Reality*, Ed. Mark Platts (London: Routledge and Kegan Paul, 1980), 69-82

Popper, Karl, *Conjectures and Refutations: The Growth of Scientific Knowledge* (New York: Harper and Row, 1963), 37-38

Posner, Richard, *The Economic Analysis of Law* (New York: Little, Brown, and Co., 1975)

Postow, Betsy, "Piper's Criteria of Theory Selection," *Southern Journal of Philosophy XXIX, Supplementary Volume: Moral Epistemology* (1990), 60 – 65

Prauss, Gerold, *Kant und das Problem der Dinge an sich* (Bonn: Bouvier Verlag, Dritte Auflage 1989)

Pritchard, H. A., "Does Moral Philosophy Rest on a Mistake?" *Mind XXI*, 81 (January 1912), 21-37

Quine, W. V. O., *Word and Object* (Cambridge, Mass.: M. I. T. Press, 1960)

_____, *Ontological Relativity and Other Essays* (New York, N. Y. Columbia University Press, 1969)

_____, *Methods of Logic*, Third Edition (New York, N. Y.: Holt, Rinehart, and Winston, 1972)

Rachels, James, *The Elements of Moral Philosophy* (New York: Random House 1986)

Ramsey, Frank P., "Truth and Probability," in *The Foundations of Mathematics and Other Logical Essays,* Ed. R. B. Braithwaite (London: Routledge and Kegan Paul, 1950), 157-198

Ranalli, Ralph, "Victims' kin decry formula for Sept. 11 compensation fund," *The Boston Globe* (January 14, 2002), A1

Raphael, D. D., Ed., *The British Moralists 1650-1800, Volume I: Hobbes-Gay*. (Oxford: The Clarendon Press, 1969)

Raphael, D. D., "Hume's Critique of Ethical Rationalism," in William B. Todd, Ed. *Hume and the Enlightenment* (Edinburgh: The University of Edinburgh Press, 1974)

Rawls, John, "Outline of a Decision Procedure for Ethics," *Philosophical Review LXVI* (1957), 177-197

_____, "Constitutional Liberty and the Concept of Justice," *Nomos VI: Justice*, Ed. C. J. Friedrich and John Chapman (New York: Atherton Press, 1963)

_____, "The Sense of Justice," *The Philosophical Review 72,* 3 (1963), 281-305

_____, "Distributive Justice," in *Philosophy, Politics and Society*, Third Series, Ed. Peter Laslett and W.G. Runciman (Oxford: Basil Blackwell, 1967)

_____, "Distributive Justice: Some Addenda," *Natural Law Forum 13* (1968), 51-71

_____, "The Justification of Civil Disobedience," in *Civil Disobedience*, Ed. H. A. Bedau (New York: Pegasus, 1969)

_____, *A Theory of Justice* (Cambridge, Mass.: Harvard University, 1971)

_____, "Reply to Alexander and Musgrave," *Quarterly Journal of Economics 88* (November 1974), 633-39

_____, "Fairness to Goodness," *The Philosophical Review 84,* 4 (October 1975), 536-554

_____, "The Independence of Moral Theory," *Proceedings of the American Philosophical Association 48,* 5 (1975; Presidential Address), 5-22

_____, "Kantian Constructivism in Moral Theory," *The Dewey Lectures 1980, The Journal of Philosophy LXXVII,* 9 (September 1980), 515-572

_____, "Social Unity and Primary Goods," in Sen, Amartya and Williams, Bernard, Eds., *Utilitarianism and Beyond* (Cambridge: Cambridge University Press, 1982)

_____, "The Basic Liberties and Their Priority," *The Tanner Lectures on Human Values, Vol. III* (Salt Lake City: The University of Utah Press, 1982

_____, "Justice as Fairness: Political not Metaphysical," *Philosophy and Public Affairs 14*, 3 (1985), 223-251

_____, "Reply to Habermas," *The Journal of Philosophy XCII*, 3 (March 1995), 132-180

_____, *Political Liberalism*, 2nd Ed. (New York: Columbia University Press, 1996)

_____, *Lectures on the History of Moral Philosophy,* Ed. Barbara Herman (Cambridge, Mass.: HarvardUniversity Press, 2000)

Raz, Joseph, *Practical Reason and Norms* (Oxford: Oxford University Press, 1990)

Real, Terrence, *I Don't Want to Talk About It: Overcoming the Secret Legacy of Male Depression* (New York: Scribner, 1997)

Reed, T. E., "Caucasion Genes in American Negroes," *Science 165* (1969), 762-768

Richardson, Henry S., "Specifying Norms as a Way to Resolve Concrete Ethical Problems," *Philosophy and Public Affairs 19*, 4 (Fall 1990), 279-310

Rifkind, Jeremy, *Time Wars* (New York: Henry Holt and Co., 1987)

Rorty, Amelie O., "Belief and Self-Deception," *Inquiry 28* (1972), 387-410

_____, Ed., *The Identities of Persons* (Berkeley: The University of California Press, 1976)

_____, Ed. *Essays on Aristotle's Ethics* (Los Angeles: University of California, 1980)

Rosendale, Don, "About Men: A Whistle-Blower," *The New York Times Magazine* (Sunday, June 7, 1987), page 56

Ross, Sir David, *The Right and the Good* (Oxford: Clarendon Press, 1968)

_____, *Foundations of Ethics* (Oxford: Clarendon Press, 1939)

Sacks, Oliver, "Neurology and the Soul," *The New York Review of Books XXXVII*, 18 (November 22, 1990), 44-50

Samkhya Karika of Isvara Krsna, trans. Swami Virupakshananda (Madras: Sri Ramakrishna Math, 1995)

Samuelson, P. A. "A Note on the Pure Theory of Consumer Behavior," *Economica 5* (1938), 61-71

_____, "A Note on the Pure Theory of Consumer Behavior: An Addendum," *Economica 5* (1938), 353-4

Savage, Leonard, *The Foundations of Statistics* (New York: Dover Publications, Inc., 1971)

Scanlon, Thomas, "Promises and Practices," *Philosophy and Public Affairs 19* (Summer 1990), 199-226

Schachtel, Ernest G., "On Memory and Childhood Amnesia," *Psychiatry 10* (1947), 1-26

Scheffler, Samuel, "Moral Independence and the Original Position," *Philosophical Studies 35*, 4 (May 1979), 397-403

_____, Unpublished comments on John Rawls, "The Basic Liberties and Their Priority," *The Tanner Lecture on Human Values*, delivered at the University of Michigan, April 1981.

_____, *The Rejection of Consequentialism* (Oxford: Clarendon Press, 1982)

Schiavo, Mary, "Flying into Trouble," *Time* (March 31, 1997), pages 52-62

Schiffer, Stephen, *Meaning* (Oxford: Oxford University Press, 1972)

_____, "A Paradox of Desire," *American Philosophical Quarterly 13* (1976), 195-203

_____, *The Things We Mean* (New York: Oxford University Press, 2003)

Schuman, Howard, Steeh, Charlotte, and Bobo, Lawrence, *Racial Attitudes in America: Trends and Interpretations* (Cambridge, Mass.: Harvard University Press, 1985)

Schwartz, Adina, "Moral Neutrality and Primary Goods," *Ethics 83* (1973), 294-307

Schwartz, Thomas, "Rationality and the Myth of the Maximum," *Nous 6* (1972), 97-117

Scitovsky, Tibor , *The Joyless Economy* (New York: Oxford University Press, 1977)

Sechehaye, Marguerite, "Excerpt from *Autobiography of a Schizophrenic Girl,*" in Bert Kaplan, Ed. *The Inner World of Mental Illness* (New York: Harper and Row, 1964)

Selby-Bigge, L. A., Ed. *The British Moralists, Vol. II* (New York: Dover, 1965)

Sen, Amartya K., *Collective Choice and Social Welfare* (San Francisco: Holden-Day, Inc., 1970)

_____, "Behavior and the Concept of Preference," *Economica 40* (1973), 241-259

_____, "Rational Fools: A Critique of the Behavioral Foundations of Economic Theory," *Philosophy and Public Affairs 6*, 4 (1977), 317-44

Shaftesbury, First Earl of, "Selections," in *The British Moralists: 1650 – 1800* (Oxford: Clarendon Press, 1969)

Shankaracharya, *Brahma Sutra Bhasya*, trans. Swami Gambhirananda (Calcutta: Advaita Ashrama, 1993)

Shapiro, David, *Autonomy and Rigid Character* (New York: Basic Books, 1979)

Shepard, R. N., "On Subjectively Optimum Selections Among Multi-Attribute Alternatives," in M. W. Shelley and G. L. Bryan, Eds. *Human Judgments and Optimality* (New York: John Wiley and Sons, 1964), 257-81.

Sibley, W. M., "The Rational Versus the Reasonable," *Philosophical Review 60* (October 1953), 554-560

Sidgwick, Henry, *The Methods of Ethics* (New York: Dover, 1966)

Simon, H. A., "A Behavioral Model of Rational Choice," *Quarterly Journal of Economics 69* (1955), 99-118

_____, "Rational Choice and the Structure of the Environment," *Psychological Review 63,* 2 (1956), 129-38

Singer, Peter, "Is Act-Utilitarianism Self-defeating?" *Philosophical Review 61*, 1 (1972), 94-104

_____, *Animal Liberation*, Second Edition (New York, NY: New York Review Books, 1990)

Slote, Michael, *Goods and Virtues* (New York: Oxford University Press, 1983)

Smart, J. J. C. and Williams, Bernard, *Utilitarianism: For and Against* (Cambridge: Cambridge University Press, 1975)

Smart, J. J. C., "An Outline of a System of Utilitarian Ethics," in Smart, J. J. C. and Williams, Bernard, *Utilitarianism: For and Against* (Cambridge: Cambridge University Press, 1975)

Smith, Holly M., "Making Moral Decisions," *Nous XXII*, 1 (March 1988), pp. 89-108.

Smith, Lillian, *Killers of the Dream* (New York: W. W. Norton & Co., 1978)

Smith, Michael, "The Humean Theory of Motivation," *Mind 96* (1987), 36-61

Sober, Elliot, "Psychologism," *Journal for the Theory of Social Behavior 8* (1978), 165-191

Spragins, Ellyn, "When The Big Paycheck Is Hers," *The New York Times* (Sunday, January 6, 2002), Section 3, 8

Stern, Lawrence, "Freedom, Blame, and Moral Community," *Journal of Philosophy* 71 (1974), 72-84

Stich, Steven P., "Could Man be an Irrational Animal?" *Synthese 64*, 1 (1985), 115-135

_____ and Nisbett, Richard E., "Justification and the Psychology of Human Reasoning," *Philosophy of Science 47* (l980), 188-202

Stocker, Michael, "The Schizophrenia of Modern Ethical Theories," *The Journal of Philosophy LXXIII*, 14 (August 12, 1976), 453-466

_____, "Desiring the Bad: An Essay in Moral Psychology," *The Journal of Philosophy LXXVI*, 12 (December 1979), 738-753

_____, "Values and Purposes: The Limits of Teleology and the Ends of Friendship," *The Journal of Philosophy LXXVIII*, 12 (December 1981), 747 – 765

_____, *Valuing Emotions* (New York: Cambridge University Press, 1996)

Strawson, P. F., *The Bounds of Sense* (London: Methuen, 1968)

_____, "Freedom and Resentment," in *Freedom and Resentment and Other Essays* (London: Methuen and Co., 1974)

Stevenson, Charles, *Ethics and Language* (New Haven: Yale University Press, 1944)

Strotz, R. H., "Myopia and Inconsistency in Dynamic Utility Maximization," *The Review of Economic Studies 23*, 3 (1955 – 1956), 165-180

Tannen, Deborah, *You Just Don't Understand: Women and Men in Conversation* (New York: William Morrow and Co., Inc., 1990)

Taylor, Charles, "Responsibility for Self," in A. O. Rorty, Ed., *The Identities of Persons* (Berkeley: The University of California Press, 1976)

Temkin, Larry, "Intransitivity and the Mere Addition Paradox," *Philosophy and Public Affairs 16*, 2 (Spring 1987), 138-187

Thomson, Judith J. and Dworkin, Gerald, Eds. *Ethics* (New York: Harper and Row, 1968), 48-70

Thomson, Judith J., *The Realm of Right* (Cambridge, Mass.: Harvard University Press, 1990)

Thurstone, L. L., "The Indifference Function," *Journal of Social Psychology 2* (1931), 139-167

Tuck, Richard, "Is there a free-rider problem, and if so, what is it?" in Ross Harrison, Ed. *Rational Action* (New York: Cambridge University Press, 1979), 147-156

Tversky, Amos, "Intransitivity of Preferences," *Psychological Review 76*, 1 (1969), 31-48

_____ and Kahneman, Daniel, "Judgment Under Uncertainty: Heuristics and Biases," *Science 185* (1974), 1124-31

_____, ""The Framing of Decisions and the Psychology of Choice," *Science 211* (1981), 453-458

Ullmann-Margalit, Edna and Morgenbesser, Sidney, "Picking and Choosing," *Social Research 44*, 4 (Winter 1977), 757-785

The Upanisads, Part I, trans. F. Max Müller (New York: Dover Publications, 1962; orig. Oxford: Clarendon Press, 1879)

The Upanisads, Part II, trans. F. Max Müller (New York: Dover Publications, 1962; orig. Oxford: Clarendon Press, 1879)

The Principal Upanisads, trans. S. Radhakrishnan (New Delhi: Harper Collins, 1996)

Upanisads, trans. Patrick Olivelle (New York: Oxford University Press, 1996)

The Upanishads: Breath of the Eternal, trans. Swami Prabhavananda and Frederick Manchester (New York: Mentor, 1964)

Eight Upanisads with the Commentary of Sankaracarya, Volume I, trans. Swami Gambhirananda (Calcutta: Advaita Ashrama, 1996)

Eight Upanisads with the Commentary of Sankaracarya, Volume II, trans. Swami Gambhirananda (Calcutta: Advaita Ashrama, 1996)

Sixty Upanisads of the Veda, Volume I, trans. from the Sanskrit Paul Deussen; trans. from the German V. M. Bedekar and G. B. Palsule (Delhi: Motilal Banarsidass, 1997)

Sixty Upanisads of the Veda, Volume II, trans. from the Sanskrit Paul Deussen; trans. from the German V. M. Bedekar and G. B. Palsule (Delhi: Motilal Banarsidass, 1997)

The Upanishads, trans. Sri Aurobindo (Pondicherry, Sri Aurobindo Ashram, 1992)

Vermazen, Bruce and Hintikka, Merrill B., Eds. *Essays on Davidson: Actions and Events* (Oxford: Clarendon Press, 1985)

Von Neumann, John and Morgenstern, Oskar, *Theory of Games and Economic Behavior* (Princeton: Princeton University Press, 1990)

Walzer, Michael, "The Obligations of Oppressed Minorities," in *Obligations: Essays on Disobedience, War and Citizenship* (Cambridge, Mass.: Harvard University Press, 1970)

Watson, Gary, "Free Agency," *The Journal of Philosophy LXXII,* 8 (April 1975), 205-220

Watson, Robert, *The Great Psychologists: From Aristotle to Freud*, Second Edition (New York: J. B. Lippincott Co., 1968

Weber, Max, *The Protestant Ethic and the Spirit of Capitalism*, Trans. Talcott Parsons (New York: Charles Scribner's Sons, 1958)

_____, *The Theory of Social and Economic Organization*, Ed. Talcott Parsons (New York: Free Press, 1964)

Wiggins, David, "Weakness of Will, Commensurability, and the Objects of Deliberation and Desire," in Amelie O. Rorty, *Essays on Aristotle's Ethics* (Los Angeles: University of California, 1980)

Williams, Bernard, "Morality and the Emotions," in *Problems of the Self* (New York: Cambridge University Press, 1973)

_____, "Egoism and Altruism," in _____

_____, "A Critique of Utilitarianism," in Smart, J. J. C. and Williams, Bernard, *Utilitarianism: For and Against* (Cambridge: Cambridge University Press, 1975)

_____, "Persons, Character and Morality," in A. O. Rorty, Ed., *The Identities of Persons* (Berkeley, Cal.: University of California Press, 1976)

_____, "Utilitarianism and Moral Self-Indulgence," in *Moral Luck* (New York: Cambridge University Press, 1981)

_____, *Ethics and the Limits of Philosophy* (Cambridge, Mass.: Harvard University Press, 1985)

Williamson, Joel, *A New People* (New York: Free Press, 1980)

Wilson, Bryan, Ed., *Rationality* (New York: Harper and Row, 1970)

Wilson, John "Freedom and Compulsion," *Mind 67* (1958), 29-60

Winch, D. M., *Analytical Welfare Economics* (Harmondsworth: Middlesex, 1971)

Winters, Barbara, "Hume on Reason," *Humes Studies V*, 1 (April 1979), 20-35

Wisdom, John, "Philosophy and Psychoanalysis, " in *Philosophy and Psychoanalysis* (Los Angeles: University of California, 1969)

_____, *Other Minds* (Los Angeles: University of California, 1965)

Wolf, Susan, "Moral Saints," *The Journal of Philosophy 79*, 8 (August 1982), 419-439

Wolff, Michael, *Die Vollständigkeit der kantischen Urteilstafel* (Frankfurt am Main: Vittorio Klostermann GmbH, 1995)

Wolff, Robert Paul, *Kant's Theory of Mental Activity* (Cambridge, Mass.: Harvard University Press, 1968)

_____, "Robert Nozick's Derivation of the Minimal State," in Jeffrey Paul, Ed. *Reading Nozick* (Totowa, NJ: Rowman and Allenheld, 1981), 77-104

Wollaston, William, *The Religion of Nature Delineated*, in Selby-Bigge, L. A., Ed. *The British Moralists, Vol. II* (New York: Dover, 1965)

Workman, P. L., Blumberg, B. S. and Cooper, A. J., "Selection, Gene Migration and Polymorphic Stability in a U. S. White and Negro Population," *American Journal of Human Genetics 15*, 4 (1963), 429-437

Wundt, Wilhelm, *Principles of Physiological Psychology*, trans. E. B. Titchener (New York: Macmillan, 1904)

Yoga Sutras: The Textbook of Yoga Psychology, trans. and commentary Rammurti S. Mishra (New York: Doubleday Anchor, 1973)

The Yoga-System of Patanjali, trans. and commentary James Haughton Woods (Delhi: Motilal Banarsidass, 1998; orig. Cambridge, Mass.: Harvard University Press, 1914)

The Yoga Sutras of Patanjali, trans. Christopher Chapple and Yogi Ananda Viraj (Delhi: Sri Satguru Publications, 1990)

The Yoga-Sutra of Patanjali, trans. Georg Feuerstein (Rochester, VT: Inner Traditions International, 1989)

The Science of Yoga: The Yoga-Sutras of Patanjali, trans. and commentary I. K. Taimni (Wheaton, Ill.: Theosophical Publishing House, 1992)

Yoga Philosophy of Patanjali, trans. from Sanskrit Swami Hariharananda Aranya; trans. into English P. N. Mukerji (Albany: State University of New York Press, 1983)

How to Know God: The Yoga Aphorisms of Patanjali, trans. and commentary by Swami Prabhavananda and Christopher Isherwood (New York: Mentor, 1969)

Patanjali's Yoga Sutras, with the Commentary of Vyasa, trans. Rama Prasada (New Delhi: Munshiram Manoharlal Publishers, 1998; orig. Allahabad: Panini Office, 1912)

Sankara on the Yoga Sutras, trans. Trevor Leggett (Delhi: Motilal Banarsidass, 1992)

Yogasutra of Patanjali with the Commentary of Vyasa, trans. Bangali Baba (Delhi: Motilal Banarsidass, 1982)

Yoga, Discipline of Freedom: The Yoga Sutra Attributed to Patanjali, trans. Barbara Stoler Miller (Los Angeles: University of California Press, 1995)

Yogavarttika of Vijnanabhiksu, Vol. I: Samadipada, trans. and commentary T. S. Rukmani (New Delhi: Munshiram Manoharlal Publishers, 1981)

Yogavarttika of Vijnanabhiksu, Vol. II: Sadhanapada, trans. and commentary T. S. Rukmani (New Delhi: Munshiram Manoharlal Publishers, 1983)

Yogavarttika of Vijnanabhiksu, Vol. III: Vibhutipada, trans. and commentary T. S. Rukmani (New Delhi: Munshiram Manoharlal Publishers, 1987)

Yogavarttika of Vijnanabhiksu, Vol. IV: Kaivalyapada, trans. and commentary T. S. Rukmani (New Delhi: Munshiram Manoharlal Publishers, 1989)

Anonymous Praise from the Referees of Cambridge University Press
for RATIONALITY AND THE STRUCTURE OF THE SELF,
VOLUME II: A KANTIAN CONCEPTION

There's much about Piper's manuscript that is interesting and good, since she is obviously extremely intelligent and knowledgeable as well as a very articulate and talented writer such that her manuscript has many significant strengths. … The manuscript's greatest strength is that it articulates a Kantian conception of the self that is interesting and has not been developed in the literature… In particular, Piper argues that certain principles of rationality (with both horizontal and vertical consistency built into subsentential concept usage) are required in order for unified agency to be possible. She argues further that this conception of agency can be linked to Kant's conception of self-consciousness in the Transcendental Deduction, contrasted with Humean desire-based conceptions, and formalized in such a way that it relates directly to "standard" versions of decision theory. She then illustrates how this theory of the self can make sense of a variety of phenomena in moral psychology. [A]nother strength of the manuscript is that it tackles an important issue, not a mere technicality, and does so without being overly narrow, as some projects in analytical philosophy can be. It is also well written and clearly organized.

… {A] highly significant contribution …. Chapter III's … main purpose … is to formulate a variable term decision calculus that modifies Savage's in certain ways. … I cannot say that there are not formal problems with the calculus that is introduced here, but I didn't see any.

Adrian M. S. Piper is a brilliant philosopher and her book, *Rationality and the Structure of the Self*, is unquestionably a groundbreaking piece of philosophical writing. The book is breathtakingly original in its attempt to explain Kant. … Piper has pushed the argument farther and better and with more subtlety than anyone writing in the area. Here I am making direct reference to neo-Kantians such as (in alphabetical order) Stephen Darwall, Barbara Herman, and Christine Korsgaard.

… [If Piper's] arguments do not succeed, then there will be none that do. Upon meeting years ago, at a conference, the now-retired Kurt Baier, I recall his speaking of the extraordinary heights of analytical abstraction and rigor that her work attained. Now that I have read this volume, I understand first-hand what Kurt Baier meant. His words did not, and could not, have done justice to the sophistication of Piper's work.

… I very much appreciated Chapter I of the book where Professor Piper discusses the pursuit of philosophy. This lends an enormous power to the book. … [I]t gives the book a kind of real-life imprimatur. To my knowledge, no theoretical book on Kantian philosophy has ever contained a chapter that speaks to the travails of being in philosophy. …

Piper's discussion of the Humean model in Chapter I is itself a model of clarity. … [her] Kantian theory of morality and the self … is theoretically subtle and elegant … magisterial … .

… Piper has made the case for the theoretical pay-ff more thoroughly and powerfully and with more subtlety than anyone in recent times has done. To my mind, no one comes even close. Indeed, in comparison to Korsgaard's argument in this regard, as given in her Tanner Lectures (for example), Piper's argument is on an entirely different and higher plane in terms of its richness and majesty and thoroughness. At the theoretical level, insofar as any one has a claim to leaving no stone unturned, Piper most certainly does.

… Chapter III, "The Concept of a Genuine Preference," is surely a *tour de force*. … I read the chapter and was extremely impressed by it. …

Piper's argument on friendship and impartiality is one of the most one of the most powerful and, if you will, morally beautiful arguments in the book. No one has made more sense of Kantian moral philosophy and the subject of friendship than she has. The very idea that friendship requires strict impartiality is absolutely marvelous. It is so often the case that Kantians writing on the subject convey a "you-just-don't-understand" attitude towards those, like Blum, who speak to the partiality of friendship. Piper is to be commended for finally, perhaps, getting us beyond that impasse. …

Working through Professor Piper's manuscript has been a most rewarding philosophical experience. I have learned so much at every step along the way. … I have been persuaded by much that Piper has written; and I have disagreed at very points. But in each every case the disagreement has been most instructive. …

Kantians will surely love the book, though each will have her or his difference with Piper's argument. However, there is this: If at this point in time I were to recommend to a dyed-in-the-wool non-Kantian moral philosopher that she or he read one book on Kant's moral philosophy, Adrian Piper's *Rationality and the Structure of the Self* would be it.

I would put RSS2 into a group of books/papers that began more or less with Thomas Nagel's *The Possibility of Altruism*. … The idea in this tradition (and RSS2 follows this) is to appeal to Kant's work to solve the contemporary problem of finding the right way to explain and justify ethical behavior; it is only secondarily intended to be exegesis of Kant's texts. … The common target of these works is the 'desire theory of action'…, which states that all actions can be traced back to desires. … [T]he contemporary source is David Hume's *Treatise of Human Nature*. [Piper] conceives this epic intellectual battle—correctly I think—as an argument between two competing pictures of

the self. Hume's self is propelled by desires, with reason just helping out to calculate how to satisfy the most desires. For Kant, reason is not merely an instrument, but a faculty that itself produces content, something that can direct action from its own principles.

[Piper]'s novel move in the project of appealing to Kant to provide justification for a non-desire based account of human action and so morality is to take Kant seriously when he says that it is one and the same reason that creates theories of the world and directs moral behavior. So she looks to the theory of cognition in the *Critique of Pure Reason* … for clues about his picture of the self in moral life. This approach has considerable plausibility because of the distinctive shape of Kant's theory knowledge. … The crucial point for [Piper]'s purposes is that, if a would-be cognizer's encounters with the world are scatty or if he lacks a concept of himself as an on-going acquirer of information, then he will lose or perhaps never develop any sense of self. … Neither Nagel nor Korsgaard looked to Kant's own theories of cognition to buttress his rationalist approach to ethics …. O'Neill is closer to [Piper] in making some use of the views of [the first *Critique*], but she does not develop this line in great detail or in anything like the same way it is developed in RSS2.

On this basis, I believe the project to be both important and original. … The writing and organization are excellent.

The central theoretical apparatus of KSS2 is given in Chapter II, where AP introduces the notions of horizontal and vertical consistency. These are important, because she will argue that consistency is necessary to having a sufficient intellectual grasp on the world to be capable of agency. This result then has two crucial implications. The first is that intellectual self-preservation and so consistency are necessary conditions to being an agent at all. So rather than reason being a potential source of action on a par with or competing with desire, an active reason that presses constantly for consistency is revealed as a necessary condition for desire or intention themselves (thus disposing of [the desire theory of action]). The second implication is that morality arises from the effort of reason to be consistent and so to preserve the life of the self.

I'm sympathetic to [Piper]'s claim that it makes sense to talk about subsentential consistency, so that the objects of one's attitudes must be understood consistently (that is, it is not just that one's attitudes or attributions to those objects must be understood consistently). …[S]he has a plausible view about Kant's understanding of representations and judgment.

…[H]er approach to the intentionality of preferences seems plausible. … [S]he makes good use of McClennen's work on resolute choice to argue that any genuinely intentional action presupposes consistency. … [She] does a good job at characterizing how we might think about reason causing action.

Humeans will complain that she shifts the burden of proof to them to say why the intellect or reason cannot cause action, but that seems fair enough after two hundred years of the burden being given to the Kantians to show that it can. …

Can reason ground morality? [Piper] makes an interesting move on this question and one that is deeply Kantian. She starts with Kant's moral theory … as providing a description of a perfectly rational/moral agent. This is certainly the way Kant saw things. Others have noticed this in the past, but I think that [Piper] uses this recognition to much better effect. Once this ideal is in place, ordinary human behavior that falls long short of perfectly morality is explained in terms of stratagems for preserving the appearance of rational consistency (and so the life of the agent as such) in the face of recalcitrant data. There is an elegance here, as well as a strong Kantian strain, where reason is the hero of ethics and rationalization (or the misuse of reason) is the villain. Varieties of rationalizing are presented in chapters 7 and 8. I think the discussion of self-deception is interesting and plausible.
… The ms. is very long, 763 pp. + a 54 pp. bibliography by my count. Still, it is not a cumbersome read; the prose marches along in quite a compelling way.

This is an important and ambitious project, right-headed in many important regards. … The author is clearly correct about several important points, central to the present study, including the following: Contemporary philosophical thought in matters moral remains deeply and pervasively influenced by often unacknowledged Humean assumptions that frame and guide philosophical accounts of moral matters. These Humean assumptions are far less plausible and justifiable than their adherent recognize. Rather than examining or justifying these Humean assumptions, they are more often enforced as a matter of professional orthodoxy. These Humean assumptions are subject to Kant's incisive criticisms of Hume. Kant's criticisms of Hume have not yet been adequately analysed, nor (accordingly) given their proper due among contemporary moral philosophers, including even the work of some self-proclaimed Kantian moral theorists, such as the late John Rawls. Kant's criticisms of Hume pertaining to moral philosophy require and deserve the extended and detailed treatment the author gives them. Contemporary moral philosophers are committed, in theory and in practice, to high standards of rational justification, although these standards and both their theory and their practice cannot be accounted for on the basis of the basically Humean views advocated by a broad swath of contemporary, especially analytic moral philosophers.…

[D]ecision theory has allowed philosophers to let their neo-Humean

presumptions run riot without critical examination. The author is right that decision theory both reflects and is central to the hegemony of neo-Humean views in moral philosophy, and that a critique of decision theory is thus a crucial part of her project. I find that the author makes judicious criticisms of contemporary decision theories and that these criticisms contribute significantly both to explaining and to justifying her Kantian alternative to neo-Humean decision theories. I also find the author's discussion of these issues deft and concise …

[T]he author's project is bold, ambitious and very timely. A great virtue of the project is to link its topic directly with issues about our own philosophical ideals and practices. The links are genuine, though the discussion of these issues about the norms governing rational discourse, especially within philosophy, has been rather scarce, especially among philosophers. … The book promises to be a block-buster … The two contrasting accounts of Moore, by Hampshire and by Keynes, are important and illuminating. … The author's discussion of various dialogical pathologies in the field only touches the tip of a very important iceberg. … the author is clearly in command of her issues, topics, and order of discussion. The discussion is all carefully and clearly formulated … The author rightly interprets Kant's moral writings within the context of Kant's *Critique of Pure Reason*, and is right that failure to do this has had wide-spread debilitating effects on 'Kantian' moral theory among analytic philosophers. … The author's objection to Brandom's inferentialist program is sound … Brandom's inferentialism is fatally flawed for just the reason the author points out.

Adrian Margaret Smith Piper (b. 1948) is an analytic philosopher and first-generation Conceptual artist. She did a B.A. in Philosophy at CCNY, an M.A. and Ph.D. in Philosophy at Harvard with John Rawls, and studied Kant and Hegel with Dieter Henrich at the University of Heidelberg. The first tenured African American woman professor in the field of academic philosophy, Piper taught philosophy at Georgetown, Harvard, Michigan, Stanford, UCSD and Wellesley. In 2012, on her 64th birthday, Piper publicly retired from being black. She has studied and practiced yoga since 1965. She lives in Berlin, where she runs the APRA Foundation Berlin and edits *The Berlin Journal of Philosophy*.