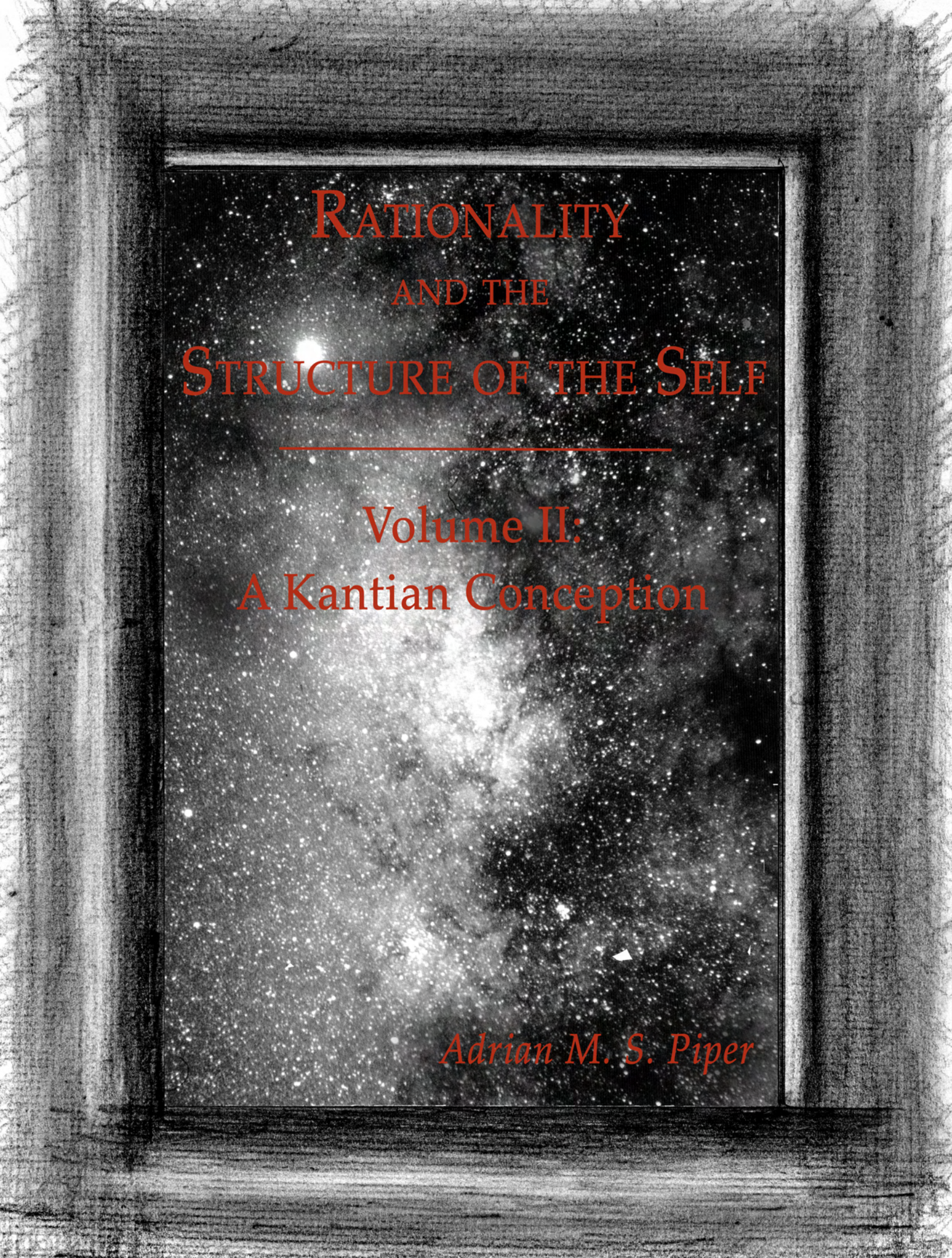
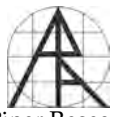

Rationality
and the
Structure of the Self
Volume II:
A Kantian Conception
Adrian M. S. Piper

RATIONALITY AND THE STRUCTURE OF THE SELF

Volume II: A Kantian Conception

Adrian M. S. Piper



First published in 2008

ISBN # 978-3-9813763-1-9

Rationality and the Structure of the Self,

Volume II: A Kantian Conception

I require of a critique of pure practical reason that when it is completed, we must be able to show its unity with the speculative in a common principle, because in the end there can be only one and the same reason, which must be differentiated solely in its application. [G, Ak.391]

To the Memory of Philip Zohn

Contents of Volume II: A Kantian Conception

List of Figures

Acknowledgements to Volume II

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' 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' 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).

Earlier versions of Chapter II were delivered to the Association for the Philosophy of the Unconscious at the American Philosophical Association Eastern Division Convention in December 1986, Akeel Bilgrami commenting; the University of Minnesota Philosophy Department in November 1987; the Columbia University Philosophy Department in March 1988; and the "Moral Psychology and Moral Identity" Conference at Oberlin College in April 1995, Michael Stocker commenting. The present version has benefited greatly from audience comments and questions received on those occasions, and particularly from those of Akeel Bilgrami, Dick Boyd, Norman Dahl, Jay Garfield, Henry Mandel, Charles Parsons, Thomas Pogge, Michael Stocker and Joan Weiner, with whom I discussed at length an early draft in 1994.

Chapter III has benefited greatly from my conversations with David Auerbach, Mark Kaplan, Glenn Loury, Ned McClennen, Robert Rubinowitz and Robert Paul Wolff, and from Kaplan's and Wolff's comments on earlier drafts. Joan Weiner provided valuable feedback when I was making final revisions. Chapter IV and parts of Chapter V were solitary but pleasurable undertakings for which I take full responsibility. Other parts of Chapter V, as well as Chapter IX, have received a good deal of exposure, for all of which I am grateful. Chapters V.5.1-2 and IX were excerpted in "The Meaning of 'Ought' and the Loss of Innocence," delivered at *The Personal Turn in Ethics* Conference at the University of Minnesota in April 1987; to the Philosophy Departments at Vassar College in October 1987, the University of Mississippi at Oxford in November 1987, the University of California at San Diego in April 1989, the University of California, Los Angeles in April 1989, the University of Colorado at Boulder in October 1989, as an Invited Paper on Ethics delivered to the American Philosophical Association Eastern Division Convention, December 1989, abstracted in *The Proceedings of the American Philosophical Association 63*, 2 (October 1989), 53-54, at Mt. Holyoke College in September 1993, Marquette University in October 1993, Georgia State University in September 1994, Oberlin College in October 1994, at the symposium, *Diskursparadigma: Form* at the University of Vienna in June 1995, the University of Utah, Salt Lake City in November 1995, and at Scripps College of Claremont Graduate School in February 1996. Comments received from each of these audiences improved these sections immeasurably. I am particularly grateful for the comments of Annette Baier, Lawrence Blum, David Brink, Jennifer Church, Joan Copjec, Norman Dahl, Keith Donellan, Terry Eagleton, Philippa Foot, John Ladd, Robert Loudon, Ruth Barcan Marcus, Warren Quinn, Rolf Sartorius, Georg

Schollhammer, Thomas Wartenberg, and Allen Wood. In addition, Barbara Herman, Christine Korsgaard and Andrews Reath made illuminating remarks about the sections that were excerpted in my essay, "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.

Work on most of Chapter VI was supported by a Woodrow Wilson International Scholars' Fellowship in 1988-1989. An earlier version of Sections 1 – 6 was published under the title, "Impartiality, Compassion, and Modal Imagination," *Ethics 101* (July 1991), 726 – 757. Still earlier ones were delivered to the Philosophy Departments of Wellesley College in November 1989,Western Michigan University in January 1990, Purdue University and Illinois State University in March 1990, the *Impartiality* Conference at Hollins College in June 1990, and at the University of Connecticut at Storrs in December 1990. I am grateful for comments received on those occasions, and also to Owen Flanagan, Charles Griswold, Ruth Anna Putnam, and the editors of *Ethics*. An earlier version of Section 7 formed the second half of "Moral Theory and Moral Alienation," *The Journal of Philosophy LXXXIV*, 2 (February 1987), 102-118. On the topics discussed there I learned much from the comments of Akeel Bilgrami, Jeffrey Evans and members of the Philosophy Department audiences at Wayne State University in November 1985, Penn State in January 1986, Georgetown, the University of California at San Diego, North Carolina State, Wesleyan, Memphis State, and the University of Minnesota, all in February 1986.

Excerpts from Chapters VII and VIII were published under the title, "Pseudorationality," in Amelie O. Rorty and Brian McLaughlin, Eds. *Perspectives on Self-Deception* (Los Angeles: University of California, 1988). I am grateful to Rorty and McLaughlin for comments on an earlier version, and to Paul Guyer and Louis Loeb for discussion. Other parts of Chapter VIII were supported by an Andrew Mellon Post-Doctoral Fellowship at Stanford University from 1982 to 1984, and published under the title, "Two Conceptions of the Self," *Philosophical Studies 48*, 2 (September l985), 173-197, reprinted in *The Philosopher's Annual VIII* (1985), 222-246. Earlier versions were presented to the Philosophy and Anthropology Group and to the Department of Philosophy at the University of Michigan; the Departments of Philosophy at Stanford in December 1982, the University of California at Berkeley in February 1983, the University of Minnesota at Minneapolis in October 1983, and the University of Pennsylvania in March 1984. I am grateful for comments received on those occasions, and also from Michael Bratman, Jeffrey Evans, and Allan Gibbard on earlier drafts. Sections 4 through 6 of Chapter VIII were delivered under the title, "The Ideal of Agent Integrity," at the University of Wisconsin/Madison Humanities Institute Conference on Art, Philosophy and Politics in April 2002, to the Yale University Department of Philosophy in February 2003, to the University of Minnesota and Indiana University Philosophy Departments in November 2006, and in German under the title, "Das Ideal von der Integrität des Akteurs" to the Ruhr-Universität Bochum workshop, Lebenswissen – Medialisierung – Geschlecht in June 2007, as part of my tenure as Marie-Jahoda Guest Professor there. I wish to thank all audiences for their comments. I am particularly grateful to Norman Dahl, and to a young man, unknown to everyone else present and evidently on

reconnaissance from another philosophy department, for motivating me to reread Frankfurt and reformulate my criticisms of him. An earlier version of the concluding paragraphs of Section 6 appeared in "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.

An earlier version of Chapter X was published under the title, "'Seeing Things'," in the *Southern Journal of Philosophy XXIX, Supplementary Volume: Moral Epistemology* (1990), 29-60, following delivery at the Spindell Conference on Moral Epistemology at Memphis State University in October 1990, Betsy Postow commenting. Postow's comments improved this chapter considerably. I also benefited from discussion with Spindell Conference participants, and particularly David Copp, Michael DePaul and William Tollhurst. Owen Flanagan and Ruth Anna Putnam offered many helpful suggestions.

Work on Sections 2 and 3 of Chapter XI was supported by an NEH Summer Stipend in 1988 and the Woodrow Wilson International Scholars' Fellowship. These sections benefited from the comments of Anita Allen, Alison MacIntyre, John Pittman, and Kenneth Winkler. It was presented under the title, "Xenophobia and Kantian Rationalism" to the Wellesley Philosophy Department Faculty Seminar and to the Cornell University Philosophy Department in February 1992; and published under that title in *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). It was also presented at the New York University Conference, *What Does the Critique of Pure Reason Have To Do With the Pure Critique of Racism? A Look at the Work of Adrian Piper* in October 1992. I learned much from discussion of these issues with commentators Paul Boghossian and William Ruddick of the NYU Philosophy Department. A revised version was delivered under the title, "A Kantian Analysis of Xenophobia," as the Plenary Address at the VII. Symposium der Internationalen Assoziation von Philosophinnen, in Vienna, Austria in September 1995; to the New York Institute for the Humanities at New York University in March 1996; and to the Humanities Institute at SUNY Stonybrook in September 1996. Work on Section 4 of this chapter was supported by the Mellon and Woodrow Wilson Fellowships, and published under the title, "Two Kinds of Discrimination," *Yale Journal of Criticism 6*, 1 (1993), 25-74; and reprinted in *Race and Racism*, ed. Bernard Boxill (Oxford: Oxford University Press), pp. 193-237. Earlier versions were delivered to the Philosophy Department at George Washington University in November 1986, the Kennedy Institute of Ethics of Georgetown University in January 1987, to the Philosophy Departments at Howard University in October 1987, the University of Mississippi at Oxford in November 1987, the City College of New York, the University of Maryland, and the Boston Area Conference on Character and Morality in April 1988, hosted by Radcliffe and Wellesley Colleges, Nancy Sherman commenting; at the Symposium, *Feminism and Racism*, at the American Philosophical Association Eastern Division Convention, Washington, D. C. in December 1988; at Franklin and Marshall College in November 1989; Williams College, in January 1990; Western Michigan University in January 1990; and at the Conference, *Ethics and*

Racism, at Brown University in March 1990. It has benefited from discussion with those audiences, and particularly from the remarks of Nancy Sherman and Kenneth Winkler. Laurence Thomas provided extensive comments on an earlier draft. Tamas Pataki extended himself far beyond the call of duty with not only penetrating comments and criticism but also much-needed editorial help on Sections 1, 2 and 5. I am particularly grateful for his patience and forbearance.

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 Schollhammer, 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.

Abbreviated Citations to Kant's Works

1C = *Kritik der reinen Vernunft*, in *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)

All references to this work are parenthesized in the text according to the standard A/B Edition pagination.

P = *Prolegomena zu einer jeden künftigen Metaphysik, die als Wissenschft wird auftreten können*, in *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)

All references to this work are parenthesized in the text according to the Prussian/ German Academy Edition pagination.

G = *Grundlegung zur Metaphysik der Sitten*, in *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

All references to this work are parenthesized in the text according to the Prussian/ German Academy Edition pagination.

2C = *Kritik der praktischen Vernunft*, in *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)

All references to this work are parenthesized in the text according to the Prussian/German Academy Edition pagination.

R = *Die Religion innerhalb der Grenzen der bloßen Vernunft*, in *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)

All references to this work are parenthesized in the text according to the Prussian/German Academy Edition pagination.

MM = *Die Metaphysik der Sitten*, in *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)

All references to this work are parenthesized in the text according to the Prussian/German Academy Edition pagination.

L = *Logik*, in *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; first published 1800)

All references to this work are parenthesized in the text according to the Prussian/German Academy Edition pagination.

3C = *Kritik der Urtheilskraft*, in *Kant's gesammelte Schriften*, herausg. königlich Preußischen bzw. Deutschen Akademie der Wissenschaften (Berlin, 1911; New York: Walter de Gruyter), Vol. 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)

All translations that appear in this Volume are mine, and deliberately sacrifice grace to literalness.

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-calibre 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.[i]

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, fineness 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 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' 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.[ii]

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:

> 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.[iii]

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-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.[iv]

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.

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 that 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.[v]

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*.

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 do 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.[vi] 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 well-being 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 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' 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, 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' 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.[vii] 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 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[viii] 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 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[ix]

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 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.[x] 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.[xi] 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.[xii] Similarly, the Noncognitivism of Allan 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 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' 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 *übermenschliche* 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 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 which have been proposed.[xiii] 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' lectures on Kant, [xiv] 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 already begun, not only in moral philosophy, but in certain branches of cognitive psychology and social theory as well.

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.

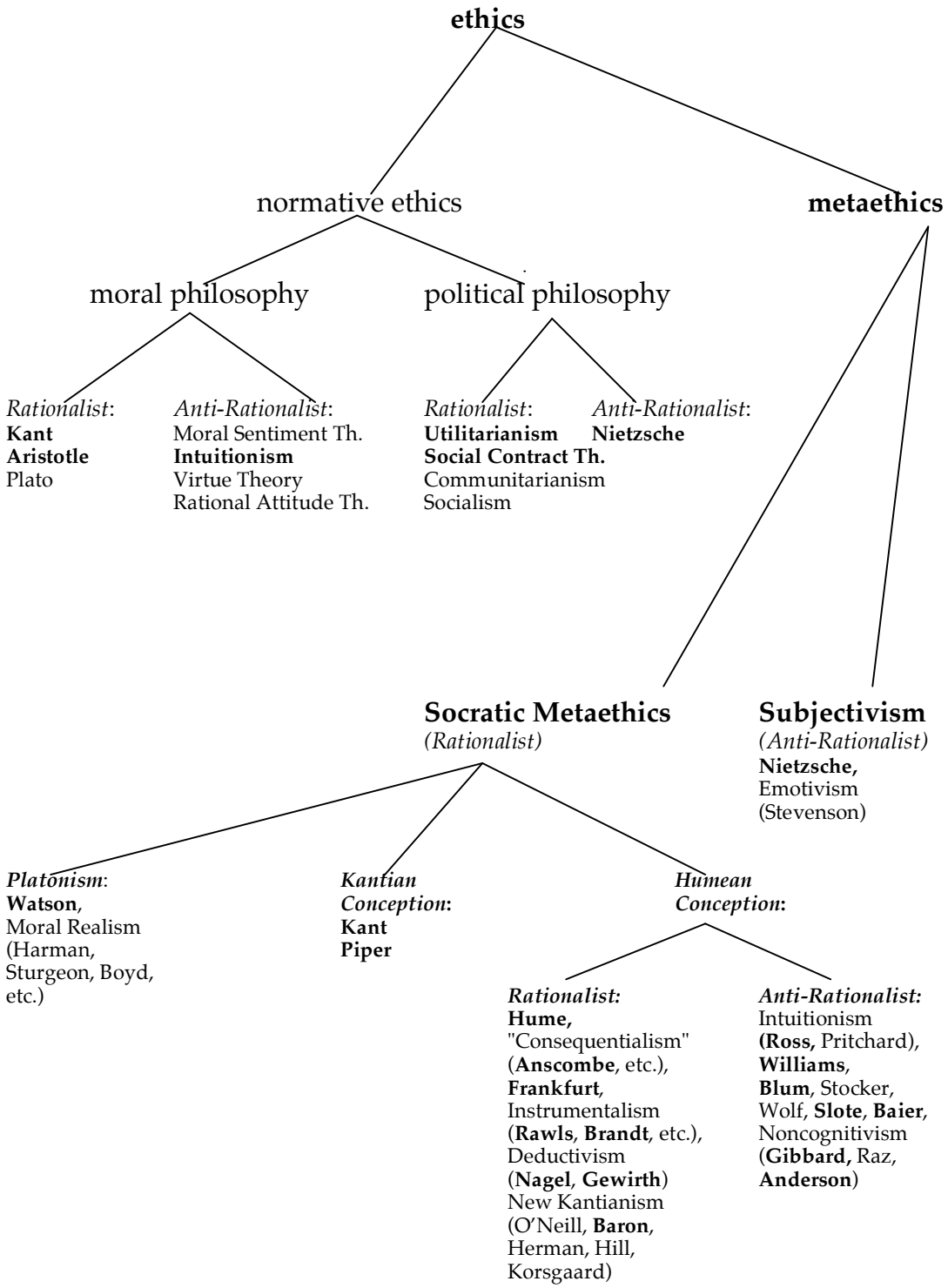
For of course Humean moral philosophers have other reasons for rationally defending their views in books and articles besides getting tenure and attracting disciples. Like Kantians, and like most philosophers, they appeal to rational argument to convince us because they believe in the rationality of 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 1. A Taxonomy of Ethics

Endnotes to Chapter I

[i]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.

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

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

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

[v]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.

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

[vii]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.

[viii]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.

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

[x]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.

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

[xii]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, 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.

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

[xiv] 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.

PART ONE: IDEALS

> Because moral laws have to be valid for every rational being as such, they are to be derived from the general concept of a rational being as such, and in this way explicate all morality, which requires anthropology for its application to human beings, at first independently as pure philosophy, that is, entirely as metaphysics …[G, Ak.412]

In Volume I of this project, I offered various arguments against the Humean conception of the self. Many of these were designed to call into question the legitimacy and veracity of the belief-desire model of motivation as a comprehensive and universally applicable account of moral motivation. These arguments did not question the obvious fact that desires sometimes do cause action. Rather, they questioned the assumption that *only* desires *can* cause actions. In this brief introduction to Part I my aim is to sketch an idealized account, not only of how reason can cause action; but of how reason always does cause action, even in those cases where desire precipitates it. That is, my aim is to outline the alternative, genuinely comprehensive and universally applicable reason-based model of motivation I develop in detail in Part I, relative to which the belief-desire model is merely a special case.

My arguments in the following five chapters are independent of and prior to any particular substantive moral theory, whether metaethical or normative, in which a Kantian conception of the self might figure. They have the same role in Kantian metaethics that the discussion in Chapters II through V of Volume I have in Humean metaethics: They aim to articulate certain foundational presuppositions in action theory, decision theory, and philosophical psychology about the nature of rationality that an adequate Kantian metaethics must presuppose. I apply this model to an account of specifically moral motivation in Chapter VI. Later chapters in Part II build on these foundations by developing criteria of adequacy for normative moral theory – without, however, attempting to make the case for one normative theory in particular over others with which it shares certain essential features in common.

In order to defend in detail this alternative, reason-based model of motivation, however, I need first to limn an alternative view of how and where in the structure of the self reason might operate. The unified account of reason I offer in Chapters II and III, following, functions simultaneously as a model of motivation and a model of rationality. It is in this sense that I claimed in Chapter I the proposed Kantian conception of the self to be simpler, prettier, and explanatorily more powerful than the prevailing Humean paradigm. Since this single, unified account of transpersonal rationality is merely an elaboration of the weak, canonical conception of theoretical reason as defined and governed by the norms of deductive and inductive logic, it is also more deeply entrenched in our thinking than the

Humean model that it is claimed instrumentally to serve. That model puts this one at the disposal of the unlimited range of contingent and variable ends particular agents adopt. Whereas none of those ends are necessary or indispensable, the canons of theoretical reason that enable their realization are both. This conception is therefore metaphysically and conceptually primary even within the Humean conception.

Whereas Hume regarded the canons of theoretical reason as mere propositional objects of calculation and computation for maximizing the satisfaction of desire, Kant maintained that the principles of theoretical reason structure the self by supplying necessary conditions for its unity. Kant thought that these principles set certain minimal requirements of logical consistency and coherence that all conscious experience must meet; and therefore that unified subjects and objects of experience must meet as well. Kant contended that any possible experience that failed to meet these requirements would be "nothing but a blind play of representations, that is, less even than a dream."(1C, A 112)

Consider the implications. In Volume I, Chapter II, I argued that the revisionist, tautological conception of a desire was not robust enough to do the needed explanatory work, and therefore was no proper desire at all. I also offered a representational analysis of desire according to which some intentional state of the agent is a desire if it includes certain sorts of conscious experience of its intentional object. A desire, on this analysis, is a certain kind of complex experience. If conformity to the minimal consistency requirements of theoretical reason is a necessary condition for integration into a unified self, and if no possible experience that fails these requirements can form part of a unified self, then in particular no possible desire that fails to conform to them can form part of a unified self. On Kant's analysis, a desire that fails the minimal consistency requirements of theoretical reason is "nothing but a blind play of representations, that is, less even than a dream." Hence no object of desire that fails these requirements can precipitate action, because no such desire can be experienced (I discuss behavior precipitated by unconscious desires in Part II, Chapters VII and VIII).

To have a desire and pursue its satisfaction in action both presuppose the existence of a unified subject whose desire and action they are. In order to have and act on a desire of any kind, then, fulfillment of the necessary conditions for unified subjecthood must be presupposed. If Kant is right in maintaining that minimal consistency requirements of theoretical reason are among these necessary conditions, then no desire that fails those requirements can motivate action because no such desire is one a unified agent can have. Motivationally effective desires as well as the final ends that are their ultimate intentional objects, then, are subordinate to the minimal consistency requirements of theoretical reason on Kant's view. I argue in Chapter II, following, that theoretical reason thus provides necessary conditions both of action and therefore of its final ends.

However, Kant's conception of the self implies even more than this. Kant's conception also implies that reason itself can precipitate action, independently of desire – and hence provides sufficient conditions of action as well. For in order that the minimal consistency requirements of theoretical reason filter out anomalous motives, emotions and thoughts from conscious unified experience, they must function effectively as sentinels, as

gatekeepers of coherence that evaluate such possible experiences for inclusion in or exclusion from conscious awareness. That is, they function as motivationally effective cognitive norms that select from the array of external and internal information and experience the *content* of both latent and occurrent thought, belief, emotion, desire, intention, and sensation, for minimal internal consistency with those which already form and constitute the structure of the self and character of the agent. Otherwise stated, unified agents are overridingly disposed to preserve their own internal rational coherence in cognitive acts of rational content-selection.

Then consider those instances in which such minimally consistent desire is absent, but occurrent thought is present, the content of which is minimally consistent relative to the agent's other experiences and dispositions, and so meets the criteria of theoretical reason. Here there need be no mystery as to what moves the agent to perform a particular action. As we saw in Volume I, Chapter VII in discussing Nagel, occurrent thoughts, beliefs, rememberings, recognizings, and so on are cognitive events with causal efficacy no less than are desires. We can distinguish among such events only on the basis of their intentional content. More specifically, I show in Chapter V below that we can distinguish the motivationally effective from the motivationally ineffective occurrent cognitive events only on the basis of their intentional content. And more specifically still, it is the intentional content of such an event that decides its degree of motivational efficacy relative to the agent's other experiences and dispositions. Most specifically of all, an agent can be motivated by the intentional content of an occurrent thought or belief to perform an action that expresses that thought or belief, whether a desire is present or not, provided that this content motivationally overrides the intentional content of other, competing beliefs and/or desires. This intentional content is rational if it satisfies the two minimal consistency conditions of theoretical reason which I develop in Chapter II, Section 4 below, and elaborate further in Chapter III, following. In Part II of this volume I describe some of the many ways in which we maintain the appearance of rational integrity even when actual rational integrity has been violated.

So, to apply this thesis to an example of specifically moral motivation for which a viable Kantian conception must provide an analysis, a whistleblower can be moved to publicize her company's unethical practices by an occurrent cognitive experience that, while minimally consistent with her other experiences and dispositions, nevertheless violates or threatens her convictions about fair labor practices. This experienced threat to her conception of fairness can motivate her to take steps to restore fairness by redressing unfairness, even though a desire to blow the whistle, whether self- or other-directed, is nowhere to be found – provided that the intentional content of this experience outcompetes in urgency the content of other intentional states she may also experience (for example, fear, self-seeking, greed, etc.). Whether it does or not depends, not on whether or not she has a "pro-attitude" toward fairness, but rather on how deeply embedded in the structure of the self the concept of fairness is for her. If it is very deeply a part of her, she will be moved to defend herself against assaults on it. Such whistleblower behavior would be a paradigm example of transpersonally rational motivation. Chapter VI.8 limns a psychological apparatus for explaining in greater detail how this could happen, and Chapter IX.8 offers a justification for why it ought to

happen. Hence the above ruminations sketch only in very general outline the argumentative strategy of this first Part of the volume. Clearly there is a great deal more to be said.

Chapter II. Reason in the Structure of the Self

In this chapter I begin to lay the foundations of my proposed solutions to the problems of moral motivation, rational final ends, and moral justification, among other problems, by showing the extent to which theoretical reason in the weak, traditional sense is a necessary condition of unified selfhood, agency, and therefore of action, in ways the Humean conception of the self fails to acknowledge. My arguments aim to remedy and revise the very superficial and purely instrumental role that the utility-maximizing model of rationality assigns to theoretical reason, and to reverse the relative priority the Humean conception of the self more generally assigns to reason and desire respectively. Whereas the Humean conception conceives desire as motivationally constitutive of the self and reason as a merely computational intellectual function for empowering it, my Kantian conception conceives reason as motivationally constitutive of the self and desire as a tangential and mostly disempowering physical impediment to its mandates.

Thus the Kantian conception I begin to develop here also rejects an approach to Kant's metaethics that views his account of rationality as an account of conscious deliberation, of which conscious computation would be an instance. Such an account would treat self-consciousness as a function whereby an agent consciously monitors the formulation and rational justification of moral maxims in light of consciously held rationality criteria. On Kant's own view, by contrast, conscious deliberation is a contingent empirical process, whereas self-consciousness is the necessary condition of conceptually unified and coherent consciousness, a condition that conscious deliberation presupposes. Kant means to supply an analysis – a "transcendental" analysis, or "rational reconstruction," if you will – of the underlying structure of moral judgment and reasoning. Right and wrong actions can be dissected as instantiating moral maxims, which in turn instantiate more general moral principles we may interrogate, independent of the contingent conscious deliberations in which a particular agent may or may not engage. For Kant, self-consciousness is a necessary precondition for unified moral agency, not a contingent product of it. My account follows Kant's in this regard.

The main thesis of this chapter is that without satisfying at least two familiar and very weak consistency requirements of theoretical reason that are deeply embedded in the structure of a unified self, we could not be motivationally effective agents at all. Here again I follow Kant in anchoring the following analysis in the principle of noncontradiction:

> Whatever be the content of our knowledge, and however it may relate to the object, the universal though merely negative condition of all our judgments in general is that they not contradict themselves; [1C, A 150] …The proposition that no predicate contradictory of a thing befits it, … holds of knowledge, merely as knowledge in general, irrespective of content; and says that the contradiction completely cancels and invalidates it. [1C, A 151]

Section 1 clears the way for defending this thesis, by critiquing an interpretation of Kant's views about cognitive structure that appropriates them into the inferentialist/ representationalist

debate. I raise some objections to Brandom's inferentialism, and offer exegetical reasons – from a non-representationalist perspective – for resisting his claim that Kant can be understood as an inferentialist. I situate my own analysis in an interpretation of Kant that comprises both inferentialist and representationalist elements. Section 2 begins this analysis by dislodging nonsentential intentional objects from the propositional attitudes in which they are conventionally embedded. I argue that nonsentential constituents of propositions are much more intimately constitutive of the self than the sentential propositions in which they figure. This view isolates and makes nonsentential intentional objects available for logical manipulation and for reference by singular terms both within the framework of classical logic, and also within the revised decision-theoretic notation I propose in Chapter III. By describing certain kinds of theoretical inconsistency that cannot be explained by applying the law of noncontradiction only to sentential propositions, I call into question the assumption, shared by inferentialists and representationalists alike, that the minimal consistency requirements of theoretical rationality – basically observation of the law of noncontradiction – apply only to the relations among sentential propositions as the atomic and irreducible bearers of sense and meaning that it is the task of logic properly to combine.[1]

Section 3 offers a contemporary version of a Kantian model of theoretical reason. I impose on cognitively accessible things and properties in general a Kantian requirement of rational intelligibility, i.e. that we should be able to recognize them as instances of concepts that constitute our perspective; and I introduce a few basic elements of the variable term calculus to be developed in Chapter III, by way of Quine's schematized axioms of identity. Section 4 derives from the requirement of rational intelligibility two further, formal consistency requirements that a coherent agent's perspective must satisfy, i.e. horizontal and vertical consistency; and shows how they can be formalized using these basic elements. These consistency requirements amount to the demand that all things that are rationally intelligible to me at a certain moment should be internally logically consistent with one another. I compare these requirements with Kant's version of them; and show that an agent's perspective is horizontally consistent if and only if it is also vertically consistent.

Section 5 invokes these two requirements in order to explain the sense in which nonsentential intentional objects might violate the law of noncontradiction, and so the sense in which the minimal consistency requirements of theoretical reason apply much more centrally to the structure of a fully unified self than the Humean view permits. Applying the requirements of horizontal and vertical consistency to nonsentential intentional objects previews my proposal in Chapter III as to how the Boolean connectives might apply to them in a revised decision-theoretic framework. Section 6 introduces a highest-order concept that all of my experiences must instantiate in order to be rationally intelligible, and argues that satisfaction of this requirement is a necessary condition of unified agency. That experience must satisfy certain basic consistency requirements of theoretical reason, then, is a necessary condition of unified agency as well.

Finally Section 7 responds to some questions and objections to this view, applies it to some extended empirical examples, and locates it relative to Thomas Nagel's analysis of double vision.

1. Is Kant an Inferentialist?

Under Hume's influence,[2] we tend to think of theoretical reasoning as a contingent set of mental operations, conscious or unconscious, that we perform on sentential propositions.[3] Call this the *Humean view of theoretical reason*, or HVTR for short. HVTR is implicit in the utility-maximization model of rationality examined in Volume I, Chapters III and IV. I make it explicit here in order to scrutinize it more closely. Whether or not we perform such contingent mental operations on sentential propositions is often thought to depend on such factors as training (e.g. whether or not we have had a course in first-order logic), personality (e.g. whether or not we are naturally "logical" or "rational" in our thinking), or the presence or absence of some object of desire we must calculate how to achieve. By conceiving of theoretical reason as a set of inferential and computational operations performed on sentential propositions, HVTR thereby situates it at a considerable remove from the kinds of factors – emotions, dispositions, desires, and so forth – we ordinarily recognize as capable of causal efficacy. It thus forecloses in advance the possibility that theoretical reason might be motivationally effective in behavior. At least it is difficult to imagine how anything so seemingly remote from causation could be.

1.1. Brandom's Inferentialism

HVTR finds support in contemporary philosophy of language under the rubric of *inferentialism*, the decompositional, "top-down" view that sentences are the primitive carriers of semantic content, from which their embedded grammatical components derive their meaning. Sentences, in turn, derive theirs from their inferential and pragmatic interrelationships in the language that generates them. These inferential relationships among sentences thus determine the referential relations of their singular terms and predicates. This holistic (actually coherence) view of language is inspired by the later Wittgenstein and by Quine, and is ably defended by Robert Brandom.[4] On Brandom's view, the primary challenge is explain in what sense, if sentences are semantically primitive, the meanings of the subsentential expressions that constitute them can be understood. "How," he asks, "can a broadly inferential approach to semantic content be extended from the grammatical category of sentences, the only sort of expression directly involved in inference, to various subsentential categories such as singular terms and predicates?" (MIE 335)

This challenge issues from *representationalism*, the compositional, "bottom-up" view that singular terms and predicates of sentences derive their meaning from the states of affairs in the world to which they refer; that sentences derive theirs from their components; and that a language comprises the sentences it has the capacity to generate. This contemporary correspondence theory of truth derives from Tarski, and its leading proponent is Jerry Fodor.[5] On this view, objects, properties and relations rather than propositional states of affairs are what are

primarily represented. Representationalism already has a plausible set-theoretic account of the way in which singular terms and predicates combine to yield sentences: Basically it assigns concrete particulars to atomic singular terms and sets of those particulars to atomic predicates, thereby generating semantic content for the sentence that comprises them. For this reason representationalism views its primary task as explaining the relationship between such representations and the cognitive representation-events of subjects.

Brandom is right to suggest that the metaphysics of the two views are not necessarily incompatible (MIE xxii, 337), and he describes two views that permit of both decompositional and compositional explanations of semantic content. He argues that, by contrast with a strict Tarskian approach, a Fregean one permits derivation in either direction: Thus, for example, the category of single-placed predicates can be derived from singular terms (T) and sentences (S) as basic categories by defining such predicates as (T → S) and stipulating that these consist in an expression that combines with a term to yield a sentence, as the expression "writes" combines with the term "Frege" to yield the sentence "Frege writes." (MIE 360-362) This derivation of single-placed predicates would be one instance of a general rule that defines any derived category (X → Y) as "a function taking arguments of the kind semantically associated with the category X into values of the kind semantically associated with the category Y." (MIE 362) This, of course, makes the question of which category is primitive and which is derivative arbitrarily dependent on which sort of entities one chooses to define as primitive in one's assignment of expressions to letters and to connectives. It does not address the issue of which in fact might be primitive in any nonarbitrary, foundational sense. But similarly, Brandom observes that the Tarskian apparatus is equally indifferent between compositional and decompositional methodologies: In Davidson's hands, truth conditions are assigned to sentences according to the beliefs and desires they express, in such a way as to make rational the explanation and prediction of the speaker's behavior. Then there should be no independent constraints on the assignment of denotations to subsentential expressions that might conflict with or pre-empt the assignment of rational-making truth conditions to the sentences in which these expressions appear. The assumption that singular terms and predicates have denotations does not commit Davidson to their primacy. (MIE 364)

If neither decompositional nor compositional methods for explaining semantic content commit one to the primacy of the type of expression stipulated to be primitive, what would? What type of explanation would ground the stipulated primitive expression in more fundamental considerations that did not reduce to arbitrary notational manipulation? For Davidson, the most fundamental consideration would be obedience to the principle of charity. By contrast, Brandom's analysis of sentential propositions as having semantic primacy aims for "the grounding and illumination of representational tropes secured by displaying the implicit features of discursive practice that are expressed explicitly by their use." (MIE xxii) That is, he aims for a demonstration that representations of things are grounded in the sentences that embed them,

rather than the other way around. The semantic primacy of sentences, in turn, is grounded in the pragmatics of their normative social use.

Brandom's strategy is first to offer an inferential explanation of the semantic content of subsentential expressions that are themselves sentences and constituent parts of compound, multi-sentential expressions; and then to derive from it a related form of explanation of the semantic content of strictly subsentential expressions such as singular terms or predicates. Brandom invokes Dummett's account of Frege's distinction between force or "freestanding sense" – what a sentential assertion commits the speaker to inferentially, and content or "ingredient sense" – how the sentential components of a compound, multi-sentential assertion contribute to the semantic content of the assertion itself. But it is notable that, as Brandom indirectly acknowledges (MIE 341), Dummett himself conceives ingredient sentences in the standard way, as having a bottom-up role in the compound sentences in which they appear:

> [S]entences may also occur as constituent parts of other sentences, and in this connection, may have a semantic role in *helping to determine the [content] of the whole sentence*: so here we shall be concerned with whatever notion of [content] is required to explain how *the [content] of a complex sentence is determined from that of its components*.[6]

Brandom proposes instead that the preservation of a compound sentence's force through substitution of one of its sentential constituents serve as a tool for understanding the semantic or ingredient content of that constituent. Two sentences have the same such content if and only if substituting one for the other preserves the force of the compound sentence in which one is a constituent. (MIE 341)[7] Similarly, two sentential propositions have the same force, or inferential content, if and only if substitution of an instance of the one for an instance of the other "never turns a good inference into one that is not good, no matter whether the sentence appears as a premise or as part of the conclusion of the inference." (MIE 347) Conjointly, these two conditions impose two requirements on substitution of a subsentential expression that is itself a sentence: first, it must preserve semantic content, or ingredient sense; and second, it must preserve force, or freestanding sense. For Brandom the latter determines what uttering the sentence commits one to; and this, in turn determines the former, i.e. what the utterance means. (MIE 348, 353)

Brandom then uses this Fregean principle of semantic invariance under substitution as a platform to launch a decompositional methodology based on the reasoning that, just as content-preserving substitution in multi-sentential inferences enables us to fix the conceptual content of single sentences, and just as content-preserving substitution in freestanding compound sentences enables us to fix the sentential content of its sentential ingredients, similarly content-preserving substitution in a simple sentence enables us to fix the content of the singular terms and predicates that are its strictly subsentential components:

> This same substitutional path that leads from inference to sentential conceptual content leads as well from the possession of freestanding inferential content by compound sentences to the possession of component-inferential content by embedded ingredient sentences and, … from sentential content to the content of subsentential expressions.

> (MIE 354) …Once this sort of ingredient content has been introduced into one's semantic theory, however, it becomes available to be associated also with expressions that (unlike sentences) can occur only as parts of assertible sentences […] such as singular terms and predicates, to which the concept of freestanding content does not apply. (MIE 359)

Following this line of reasoning, strictly subsentential categories of linguistic expression can be defined using Frege's notion of substitutional invariance: two strictly subsentential expressions are of the *same grammatical category* if and only if substituting one for the other preserves the sentential status of the well-formed sentence in which one of them occurs. Two strictly subsentential expressions have the *same semantic content* if and only if substituting one for the other preserves the "pragmatic potential" – that is, the inferential force of the sentence in which one of them occurs. (MIE 368)

Singular terms are then distinguished from predicates by the directionality of the substitution inferences which substitutional invariance yields. Substituting one singular term for another with the same semantic content in a sentence yields a *symmetric* inference from the truth of the original sentence to the truth of the sentence containing the substituted term: "Benjamin Franklin invented bifocals" is true if and only if "The first postmaster general of the United States invented bifocals" is also true. By contrast, substituting one predicate for another with the same semantic content in a sentence yields an *asymmetric* inference from the truth of the first to the truth of the second: If "Benjamin Franklin walked" is true, then "Benjamin Franklin moved" is also true; but not vice versa. So singular terms with the same semantic content satisfy equivalence relative to the substitution inferences they yield, whereas predicates need not. (MIE 372)

Although I am sympathetic to Brandom's inferentialist program, I do not think it shows that sentential propositions, or sentences, are the primitive carriers of semantic content. At most it shows that sentential propositions can be construed in this way. But in order to show that they really are semantically primitive, or primary, Brandom must ground this construal in more fundamental considerations that go beyond the bidirectional explanatory flexibility that, as he has acknowledged, both decompositional and compositional methodologies allow. The pragmatics of normative linguistic usage are the more fundamental considerations that Brandom offers to anchor his decompositional analysis. However, in order for the pragmatics of normative linguistic usage to function in this way – i.e. to have explanatory import over and beyond Brandom's decompositional analysis itself, they must correspond (you will pardon the term) to the actual norms according to which speakers use sentences, predicates and singular terms. And this desideratum comes into collision with his deployment of the Fregean principle of substitutional invariance for fixing the semantic content of strictly subsentential expressions.

This principle succeeds in demonstrating when two sentences with different singular terms have the same semantic content, but it does not provide a criterion for identifying those which do. The principle presupposes that we already know when two singular terms have the same semantic content and when they do not. In order to make use of the principle of substitutional invariance between sentences, we first need to know which singular terms are

mutually equivalent such that sameness of semantic content between sentences is preserved. Unless we first know that "Benjamin Franklin invented bifocals" is true whereas "Clark Kent invented bifocals" is false, *and why*, namely that "Benjamin Franklin" and "Clark Kent" do not denote the same concrete particular, there is no way for us to determine whether intersubstitution of these two singular terms in the respective sentences preserves semantic content or not – nor, therefore, whether using the respective sentences interchangeably preserves pragmatic force or not. Actual linguistic application of the Fregean principle requires that the semantic content of strictly subsentential expressions have been fixed in advance. And this, in turn, argues in favor of ascribing a causally and epistemically primitive role to those strictly subsentential expressions.

Yet Brandom barely considers the possibility that the causal and epistemic primacy of singular terms in early acculturation and subliterate psychological processes such as dreaming and fantasizing might suggest an answer to the question of grammatical primacy:

> It is one thing to claim (how could it be denied?) that causal interactions of various sorts with particular objects is a necessary condition of being able to represent empirical states of affairs; it is another to claim that some of these interactions ought to be understood as semantically primitive, in that what it is to represent such a states of affairs ought to be understood in terms of them. (MIE 337, fn. 2)

Brandom is of course right to distinguish these two claims. But having acknowledged the import of the first, he does not say why it does not endorse an inference to the second; or why, therefore, the second does not provide a straightforward answer to the deeper question of which hierarchical order best respects the cognitive facts about whether it is sentences or strictly subsentential expressions that have primacy and intimacy in the structure of the self. Without some such causal relationship between things and the singular terms by which we learn to denote them, it is hard to see how truly pragmatic and functional norms of linguistic usage could develop.

1.2. Brandom's Kant

Brandom would react with dismay to my suggestion that his brand of inferentialism supports HVTR, for he takes himself to be a good Kantian (as we have seen in Volume I, this reaction would not be unusual among the many Humeans who take themselves to be good Kantians). In fact he appeals to Kant's authority in defending the primacy of sentential propositions; but I am not convinced by this appeal, either. Brandom contrasts Kant's view with what he calls the "pre-Kantian tradition," according to which

> (A) The proper order of semantic explanation begins with a doctrine of *concepts* or *terms*, divided into singular and general, whose meaningfulness can be grasped independently of and prior to the meaningfulness of judgments. Appealing to this basic level of interpretation, a doctrine of *judgments* then explains the combination of concepts into judgments, and how the correctness of the resulting judgments depends on what is

> combined and how. Appealing to this derived interpretation of judgments, a doctrine of *consequences* finally explains the combination of judgments into inferences, and how the correctness of inferences depends on what is combined and how. (MIE 79)

Kant, Brandom claims, rejects this pre-Kantian tradition; and offers, as "one of his cardinal innovations," the thesis that the judgment is the "fundamental unit of awareness or cognition, the minimum graspable." (MIE 79) In support of this claim Brandom quotes Kant's assertion at 1C, A 69/B 94 that all acts of the understanding can be reduced to judgments, that the understanding is the faculty of judging, and that concepts can be used by the understanding only to form judgments. From this passage Brandom concludes that "for Kant, any discussion of content must start with the contents of judgments, since anything else only has content insofar as it contributes to the contents of judgments." (MIE 80)

Actually this conclusion is a bit too strong to represent the full complexity of Brandom's interpretation of Kant, since he has previously characterized Kant's view of concepts as one according to which they have the form of rules, and hence specify "how something *ought* (according to the rule) to be done." (MIE 8) If concepts themselves specify how something ought to be done, then they have content independent of their role in judgment. Understanding, according to Brandom's Kant, is the conceptual faculty of grasping rules – "of appreciating the distinction between correct and incorrect application they determine." (*ibid*.) If understanding grasps the rules that constitute the concepts they form, then the understanding grasps content that its concepts already have; and those concepts themselves, rather than the judgments in which they figure, must be the "minimum graspable."

Moreover, Brandom's Kant accepts the rationalistic, classificatory account of cognition, according to which intuitions are classified under concepts, against empiricist claims that not all awareness presupposes conceptual classification:

> (B) All awareness is understood as exhibiting the classificatory structure of universal or repeatable concepts subsuming particulars. … Kant denies apprehension without classification, insisting that there must be conceptual classification wherever there is any sort of awareness. Awareness of what is classified and of how things can be classified derives from awareness that consists in classifying. (MIE 86)

This is a fairly accurate gloss on Kant's insistence on the necessity of classification – or conceptualization – for conscious experience. If, as Brandom asserts, awareness for Kant is necessarily conceptual awareness, and conceptual awareness consists in classifying and subsuming particulars according rules that specify how these particulars ought to be classified, then concepts, not judgments, are "the fundamental unit of awareness or cognition, the minimum graspable." Thus Brandom's Kant does, after all, ascribe content, awareness, and minimum graspability to concepts independently of their role in judgment.

Brandom interprets Kant's notion of necessity to mean "'in accord with a rule,'" (MIE 10) and hence to imply the "necessity" of conceptual specifications of how something ought to be done. He distances Kant's conception of necessity from that of contemporary discussions of

modality on the grounds that Kant's concerns are fundamentally normative and practical rather than descriptive and theoretical (*ibid*.) Presumably Brandom's Kant would not find necessity in the conformity to just any rule conceptually specifying how just anything ought to be done. For example, it is not likely that there would be any necessity in the rule that specified that one's teeth ought to be brushed back to front rather than front to back. Nor would one expect to find necessity in the rule that predicates fiscal transparency of corporate accounting offices. Presumably only certain kinds of conceptual specifications of how certain kinds of things ought to be done have necessity in this sense. But if necessity just is conformity to a rule, as Brandom's Kant claims, then the necessity that distinguishes these particular conceptual specifications consists, presumably, in according with some further rule that conceptually specifies how these particular conceptual specifications ought to function; and the necessity of this rule, in turn, in according with yet a further one that conceptually specifies its functioning. Hence either the sense in which any particular conceptual specification is necessary is always at one remove from the conceptual specification itself; or else Brandom's Kant cannot mean to identify necessity with conforming to a rule simpliciter. There has to be more to necessity than this.

Brandom further characterizes Kant's faculty of understanding as the "active, cognitive faculty" that "synthesize[s], bring[s] things into a unity – that is, subject[s] them to rules or concepts." (MIE 80) That synthesizing activity, he asserts, "is an aspect of judging." In support of this assertion he quotes Kant's own claim at 1C, A 79/B 104 – 105:

> (C) (1) The same function which imparts unity to various representations in one judgment (2) imparts unity likewise to the mere synthesis of various representations in one intuition, (3) which in a general way may be called the pure concept of the understanding. (4) The same understanding, and by the same operations by which in concepts it achieves through analytical unity the logical form of a judgment, (5) introduces also, through the synthetical unity of the manifold in intuition, a transcendental element into its representations.

In his footnote to this citation, Brandom adds that "the 'transcendental element' introduced in this way is just reference to objects." (MIE 80 fn. 18) I discuss this passage below.

1.3. My Kant

I do not agree with Brandom that Kant rejects the "pre-Kantian tradition" described in passage (A) above. However, I also do not think that Kant accepts it – at least not in this form. Nor do I agree that 1C, A 69/B 94 shows that Kant believed the judgment to be "the fundamental unit of awareness or cognition, the minimum graspable." (MIE 79) Above I offered some evidence that Brandom does not entirely believe this, either. Finally, I do not think Brandom is justified in appealing to Kant's authority in support of his inferentialist program. However, I also do not think this makes Kant a representationalist. Kant's view is a more complex one that incorporates signature elements of both views. A full defense of these opinions is unnecessary for purposes of this discussion. I undertake only as much of one as I think necessary in order to

anchor my own discussion of subsentential expressions, in subsequent sections of this chapter, in my own understanding of Kant's view. So I shall largely confine my remarks here to further examination of passage (C), above, which is from Kant's introduction to the Table of Categories.

Brandom makes some significant translation choices and edits to passage (C). The original runs as follows:

> (C') (1) *Dieselbe Funktion, welche den verschiedenen Vorstellungen in einem Urteile Einheit gibt,* (2) *die gibt auch der bloßen Synthesis verschiedene[r] Vorstellungen in einer Anschauung Einheit,* (3) *welche, allgemein ausgedrückt, der reine Verstandesbegriff heißt.* (4) *Derselbe Verstand also, und zwar durch eben dieselben Handlungen, wodurch er in Begriffen, vermittelst der analytischen Einheit, die logische Form eines Urteils zustand brachte,* (5) *bringt auch, vermittelst der synthetischen Einheit des Mannigfaltigen in der Anschauung überhaupt, in seine Vorstellungen einen transzendentalen Inhalt,* (6) *weswegen sie reine Verstandesbegriffe heißen, die a priori auf Objekte gehen,* (7) *welches die allgemeine Logik nicht leisten kann.*

In this passage Kant twice deploys what he in the *Prolegomena* calls the "analytic" or "regressive method," [P, Ak. 264, 274, 276 fn] of beginning with the empirical fact of judging and working backward to its necessary preconditions: first in the transition from (C'.1) to (C'.2); and second in the transition from (C'.4) to (C'.5). Let us take each numbered phrase in turn.

(C.1) is a straightforward translation of (C'.1). (C.2) is not quite a straightforward translation of (C'.2), because the primary meaning of *bloß* is "bare" or "naked," not "mere." By modifying the noun "synthesis" with the adjective "bare," Kant means to call attention to the distinction between the unmediated and unadorned cognitive operation of gathering diverse representations together simpliciter, and the higher-level operation of giving them cognitive unity. For this it is not sufficient that the representations simply land, as it were, in a heap in inner sense. In order to achieve cognitive unity, the representations must be gathered and sorted according to an organizing principle that the concept under which they are gathered supplies. Hence (C'.1) plus (C'2.) together say that there is one function that does two things. It unifies various representations into one judgment. It also unifies the bare synthesis of various representations into one intuition – a necessary condition for judgment.

Kant says at 1C, A 19/B 33 that, regardless of the kind and means by which cognition relates to objects, intuition is in unmediated relation to them; and that all thought is directed at intuition. Hence all thought is directed at our conceptually unmediated relation to objects. And at 1C, A 68/B 93 Kant defines a "function" as the unity of the act of ordering various representations under one common representation (Kant uses the term "representation" to refer to any mental contents [1C, A 320/B 376], so we must rely on context to establish the metaphysical level and kind of representation he means to denote). So in order for this function to unify representations in a judgment, it first must have unified the bare – unmediated – synthesis of representations in an intuition. The synthetically unified representations that constitute an intuition are then unified, along with other such intuitions, in a judgment. Hence intuitions are metaphysically prior to judgments, and the representations that synthetically

constitute intuitions are metaphysically prior to the intuitions themselves. Kant asserts this explicitly at 1C, B 145. – This is the kind of assertion that might lead the unsuspecting to think that Kant is a representationalist.

(C.3) is a straightforward translation of (C′.3): this double-barreled cognitive function is called the pure concept of the understanding. Thus pure concepts of the understanding have two cognitive functions for Kant, enumerated here in order of metaphysical primacy (i.e. synthetically or progressively): first they synthetically unify into intuitions the unmediated representations which we directly receive from objects; and second, they unify the mediated representations that constitute intuitions into judgments. This passage signals an important shift in explanatory strategy from that which Kant deployed in the Transcendental Aesthetic. There he seemed to want to treat intuitions as epistemically primitive and also as metaphysically independent of the higher conceptual functions of the understanding; see also 1C, A 89/B 121 – A 91/B 123. In (C′), by contrast, he leaves no room for doubt that rule-governed conceptual synthesis of diverse representations is a precondition even for a unified intuition. Hence the pure concepts of understanding are here seen to operate "all the way down" to the first moment of reception of unmediated object-representations in inner sense. – This, by contrast, is the kind of assertion that might lead the unsuspecting to think that Kant is an inferentialist.

(C.4) is a straightforward translation of (C′.4), and makes the interesting point that the categories of the understanding are metaphysically prior to the logical forms of judgment (however it should be noted that Kant reverses this order of priority at G, Ak. 454 in the *Groundwork*). Since the categories are distinguished from the logical forms of judgment by their "transcendental content," (C′.4) implies that the source and character of that content make a significant contribution to the conceptual structure of cognition. (C.5) is not a straightforward translation of (C′.5), because *Inhalt* means "content," not "element" (the correct translation of "element" in German is *Element*). Thus (C′.4) plus (C′.5) says that understanding – the same synthetic conceptual function Kant has just discussed – does two things through the very same action. By securing the analytical unity of (pure) concepts, it brings forth the logical form of a judgment. And by securing the synthetic unity of the manifold (representations) in an intuition, it introduces a transcendental content into those representations – again, a necessary condition for the analytical unity of concepts and hence of judgments.

There can be no serious question as to how closely committed Kant is to the notion of transcendental content in this passage, because on the previous page [1C, A 77 – A 78/B 103], Kant has declared that representations must first be given in order for us to analyze them; and that the content of concepts therefore cannot arise through analysis. Kant has then gone on to describe synthesis of a manifold, whether pure or empirical, as that which collects the elements [*Elemente*] into a cognition and unifies them into a particular content [*zu einem gewissen Inhalte vereinigt*]. Similarly, Kant has identified the content of knowledge with its matter at 1C, A 6/B 9 and 1C, A 59/B 83; and at 1C, A 143/B 182 in the Schematism goes on to declare that what in the

object corresponds to sensation in the subject is "the transcendental matter of all objects as things in themselves (thinghood, reality)."

Thus Kant's explanation of conceptual content runs as follows.[8] Through the process of directly intuiting objects in themselves, we receive unmediated representations from them in inner sense. We then synthesize these intuitional representations according to a certain kind of conceptual function that organizes and unifies them. By thus unifying them conceptually, we give them content. This "transcendental content" – i.e. content generated by objects to which we have no unmediated conceptual access – is the unified analytical content of the pure concepts of the understanding, i.e. those which conjointly determine how we conceive objects. This analytical conceptual content in turn provides the logical form of judgments we make about them. Now I suppose it would be possible to quibble about the distinction between transcendental content, conceptual content, and semantic content. But I do not think this would be worthwhile, because it would not obscure the most important point, that judgment is not the fundamental unit of awareness for Kant; intuitional representations are.

Moreover, judgment is not even the fundamental unit of cognition for Kant; pure concepts are (Kant distinguishes between awareness and cognition throughout the Paralogisms, but see especially 1C, A 360 and B 414 fn.). Kant in passage (C') then goes on to add that it is because of their transcendental content that such representations are called pure concepts of understanding that apply a priori to objects (C'.6), which general logic cannot do (C'.7). That is, the concepts that necessarily apply to all objects of experience do so because they gather and organize a manifold of unmediated intuitive representations from those objects in themselves. We have synthesized and unified these representations into those very concepts which conjointly define what an object is.

Thus there are certain concepts that always contain a direct and conceptually unmediated connection to the objects they denote, regardless of the particular character of those objects, namely those concepts which conjointly set the conditions something must satisfy in order to be an object of experience at all; these are the pure concepts, or categories, of the understanding. Kant enumerates these concepts in the Table of Categories at 1C, A 80/B 106. By contrast, general logic – the Table of Judgments at 1C, A 70/B 95 – cannot apply a priori to all objects of experience because they have no such content; they are mere forms of judgment. Judgment forms without content are nothing more than syntactical containers for the semantic content that denotational conceptual representations provide. Hence it is simply not true that "for Kant, any discussion of content must start with the contents of judgments, since anything else only has content insofar as it contributes to the contents of judgments." (MIE 80) At the most primitive cognitive level, things have content for Kant insofar as they contribute to the representational content of the concepts that denote them.

Again the unsuspecting might jump to the conclusion that this makes Kant a representationalist – or, to use the older term, a correspondence theorist of truth á la Tarski. I do not think it does, because Kant expresses his misgivings about such a view at 1C, A 58/B 82.

What he calls the nominal definition of truth as the agreement of knowledge with its object cannot be right, he argues, because it does not provide a general criterion of truth at all. Since each object is different, each true conceptual representation of it will have different content and a different relation to the object that makes that representation a true one. But a general criterion of truth would have to be satisfied by all such representations. Since what makes each such representation true is different in each case, no such general criterion can be given. He concludes that a criterion of truth that is both sufficient and general is impossible. Note that he is not denying that knowledge might agree with the objects it denotes. Nor is he denying that such agreement might constitute a semantic primitive in his analysis of intuition, concepts, and judgment. All he is denying is that a meaningful criterion of truth might be extracted from such agreement. If the agreement of knowledge with its object provides no leverage for a truth criterion, a fortiori it can provide no leverage for a representationalist truth criterion.

Kant thinks a coherence theory of truth, aka inferentialism, is equally insufficient. If we abstract from the content of knowledge and consider merely its form, he says, we are left with the forms of logical judgment enumerated in the Table of Judgments at 1C, A 70/B 95. These certainly do supply universal and necessary criteria of truth in the sense that whatever contradicts them must be false. But they do not establish that whatever fails to contradict them is true, because a representation that satisfies them still may be contradicted by its object [1C, A 59/B 84]. The linguistic holism that inferentialism endorses might weave a tight and complex web of sentential inferences indeed, which nevertheless bore no truth-preserving relationship to the denotations of its singular terms and predicates. If the inferential relationships mapped in the Table of Judgments provide no sufficient condition for determining whether or not a representation that satisfies them succeeds in denoting its object, then as far as Kant is concerned, it provides no leverage for a materially robust inferentialist truth criterion.

So whereas the denotational relationship offers agreement with the object but no general criterion of truth, the inferential relationship offers a general and necessary criterion of truth but no guarantee of agreement with, i.e. denotation of the object. The sufficient condition of truth that such agreement would provide cannot be stated in a general form. Hence neither is adequate, either singly or conjointly, to provide an answer to the question of what truth is. From this conclusion Kant can now argue that an answer to that question can be found only within the limited realm of empirical experience itself, in which we agree to leave investigation of the cognitive and metaphysical preconditions for having such experience out of account. While we can ascertain whether an empirical assertion is or is not true to the facts we observe, we cannot ascertain whether or not the facts as we empirically observe them are or are not true to the noumenal reality to which we futilely intend our assertions to refer.

What we have seen from close analysis of passage (C') above is that Kant's view synthesizes key elements of both inferentialism and representationalism. It is inferentialist in its defense of a restricted set of judgment forms that bear logical interrelationships and circumscribe the scope and types of judgments it is humanly possible to make. It is representationalist in

insisting on a causally direct and unmediated relationship between certain concepts that enter into such judgments, and the real world objects those concepts represent. Brandom is quite right to argue, as he does in passage (B), above, that Kant requires conceptual classification of an object as a necessary condition for experiencing it. Brandom is also right to insist that concepts must combine in the right ways in order for us to make judgments about those objects. Where he goes wrong is in thinking that we could make such judgments, and could understand their singular terms and predicates, without any independent representational relationship to the objects those subsentential expressions denote. Kant does not make that mistake because of the foundational role he accords the notion of a concept as a function for unifying representations. In the following sections I hope not to make that mistake either, and for much the same reasons.

2. Nonsentential Intentional Objects

Although both inferentialism and representationalism acknowledge the existence of strictly subsentential expressions, each assigns them a different semantic function: the first as semantically derivative from the sentences in which they are nested; the second as semantically primitive elements from which sentences are constructed. However, neither Brandom nor Fodor and Lepore acknowledge the semantic implications of taking strictly subsentential expressions as intentional objects of their own philosophical discourse. That is, all three seem to assume, along with HVTR, that intentional attitudes[9] can take only sentential propositions, and not strictly subsentential expressions, as objects. None acknowledge the existence, semantic significance or occasional syntactical intractability of nonsentential intentional objects. I now argue that they should.

2.1. Intentionality and Sententiality

Consider any proposition of the form, "I believe that P." Because *P* here can be expanded into a sentential proposition which itself may be true or false, it is natural to assume that any object of an intentional attitude can be treated similarly. But this is not so. Some intentional attitudes require a more fine-grained analysis, and thereby illuminate the overall flat-footedness of the familiar one. I focus here on intending, but intend my conclusions to have general application.

Take the sentential proposition,

(1) I intend to go to the store.

If any object of an intentional attitude itself can be expanded into a sentential proposition, we ought to be able to do so with the intentional object of (1). But how? Here is one seemingly obvious candidate:

(2) I intend that I go to the store.

Since the intentional object of (2) is itself a proposition which may be true or false, (2) fits the familiar pattern, "I intend that P."

The problem is that (1) and (2) are not semantically equivalent. I can carry out (2) by first going to a hypnotist who instills in me the motivationally effective command to go to the store, and then somnolently carrying out that command. Or I can take a pill, or get a neurological implantation, or any number of other familiar agency subverters that get me to do what the intentional object of (2) requires, namely go to the store. To intend *that* I go to the store is to intend *to* bring it about, by whatever means are available to me, that I go to the store, even if the locomotive behavior constitutive of my going to the store is not itself voluntarily undertaken. In general, this is because to intend that P is to intend that some *independent* state of affairs, expressable in a sentential proposition which itself may be true or false, obtain. Thus *that I intend to go* to the store may be false, although *that I go* to the store is true, and vice versa. To intend *that* this independent event occur is to intend *to* bring about something – my going to the store – that itself bears no necessary relation to my own agency.

This means that there is often no difference in the degree of voluntariness expressed between (2) and

(3) I intend that Clive go to the store

i.e. not much voluntariness at all. In both cases, my role may be merely *to bring it about* that the agent goes to the store – by cajoling, threatening, exhorting, hypnotizing, or implanting an electrode, without voluntarily or deliberately carrying out the object of my intention at all.

For example, suppose I know that in two hours I will have fallen asleep, and will be incapable of deliberately carrying out any sustained plan of action whatsoever; but that it is nevertheless imperative that I go to the store in two hours. I may, through autohypnosis, implant in myself the suggestion that when I hear the clock strike five, I will interrupt whatever I am doing and go to the store. At five PM I hear the clock strike five times; I awake with a start, lace up my sneakers, and stumble off to the store. My behavior is goal-directed, so it is intention*al*. But for all the direct relation it bears to my original *intention* that I go to the store, it might just as well have been Clive whom I hypnotized as myself. This is the kind of case in which "intention that" and "intention to" locutions are not interchangeable.

By contrast, I cannot carry out (1) by thus allowing hypnosis to subvert my agency. If I intend *to* go to the store, then whatever means I deploy to do so (a pair of sensible shoes, a bus, etc.) cannot involve putting someone or something else in direct command of my will in order to do so. In general, this is because to intend to do something requires that the event I intend bear a necessary relation to my own agency, i.e. that it be not only my behavior, but moreover under my voluntary control at the time I perform it.

Note that this point is not affected by the so-called "accordion effect."[10] Even if we redescribe the intentional object of my intention as, say, getting some food in the house, it is still true that if I intend *to* get some food in the house, whatever locomotive behavior of mine is involved in doing so must be under the direct control of my will. Here I do what I do *because* I intend to do it. By contrast, if I intend merely *that* I get some food in the house, or *that* the house

be well-stocked with food, there is no reason not to call on the hypnotist (or the pharmacist, or the mad scientist) to bring this about.

Nor is the point affected by cases in which what I intend to do is effect some long-term goal to which certain instrumental actions on my part are means. Consider, for example, the case in which I intend *to* stop smoking, and achieve this *by or through* a combination of hypnosis and behavioral reconditioning. May I not say that I fulfilled my intention to stop smoking, even though most of the locomotive behavior through which I achieved this was not under the direct control of my own will? I think not. The more precise expression would be that I *resolved* to stop smoking, or *resolved at all costs* to stop smoking; and deployed these agent-independent means to achieve my *resolve*. Correlatively, to *intend at all costs* to stop smoking reveals the asymmetry: this goal can be thwarted only by subverting the intention, whereas the resolve at all costs to stop smoking can be thwarted by continuing to smoke. Thus even here, the "intend to" locution connotes an act, or series of acts of will: I would deflate my insistence that I had done what I intended to do by then allowing that I had in fact paid a hypnotist to stop me. Similarly with dieting: I would undermine my claim to have fulfilled my intention to eat less if I then admitted that I had achieved this by getting a dentist to wire my jaws shut. This is because in all such cases, I succeed in doing what I intend to do only if the actions by which I do it are under the direct and unmediated control of my will.

Is there any other reformulation that both preserves the meaning of (1) *and* is a credible candidate for being that which an agent intends to do? Consider, for example,

(4) I intend that I go to the store deliberately, by means of this very intention.

But to act deliberately does not imply that the action is under the direct and unmediated control of my will, as the Manchurian Candidate himself might remind us. The intentional object of (4) merely reiterates the same gap between action and will as does (2). Appending “this very intention” as the means in effect stipulates (2) as the means by which I achieve (2). This succeeds only in reiterating the problem at issue, by appending once more the very same gap. Or consider

(5) I intend that I go to the store, such that my going to the store occurs because I intend to go to the store.

(5) Closes the gap in (2) by stipulating (1) as its cause. But this does not show that (1) itself can be expanded into a sentential proposition. Nor does it show that (1) and (2) are equivalent; quite the contrary. It thus provides fuel for my argument, not for HVTR.

All such substitutions suffer two general defects. First, credibility: agents do not ordinarily intend, in addition to everything else, that their behavior remain under the control of their own agency, even though it must in order for them to intend to do anything. Second, intentional fidelity: such extended and philosophically complex sentential analyses of the objects of intentional attitudes run aground on the commonsense objection that if such an analysis does not happen to capture what a particular agent claims sincerely to have had in mind, then either they by definition describe a different intentional attitude, or else need to be supplemented by an argument against even *this* kind of first-person authority. This should be kept in mind in the

treatment of (9), below. Then if (1) and (2) are not semantically equivalent, not all objects of intentional attitudes themselves can be reformulated as sentential propositions. Call those that cannot *nonsentential* intentional objects.

My first proposal may be put as follows:

> *Proposal 1:* Anything that may occupy the subject or predicate position in a sentential proposition that does *not* express an intentional attitude, such as, for example,
>
> (6) To go to the store is a tedious errand
>
> also may be a nonsentential intentional object, as is the subject of (6), "to go to the store," in (1).

Here are some other examples of nonsentential intentional objects that singular terms might more conventionally denote: "the number 3" in

(7) I am thinking of the number 3;

"the situation in Africa" in

(8) I am thinking of the situation in Africa;

and so on. Like the intentional object of (1), the intentional objects of (7) and (8) cannot be reformulated as sentential propositions, because they do not ascribe properties to anything. Rather, they themselves may correspond to properties, or to the events, particulars or states of affairs to which properties are ascribed. It would seem that there are many such nonsentential intentional objects. In fact, anything we can think of (literally), i.e. *any concept we have*, of going to the store, the number 3, the color purple, Vienna, the situation in Africa, and just about any other, *may function as a nonsentential intentional object* in a sentential proposition of the form, "I am thinking of ...."

Other intentional attitudes, like that of intending to do something, are more restrictive in the range of objects they may take; but not because all such objects must be sentential. Indeed, a tentative inductive generalization may be in order: It is a rare intentional attitude indeed that takes intentional objects *none* of which resist reformulation as sentential propositions.

2.2. The Psychological Primacy of Nonsentential Intentional Objects

It may seem that all nonsentential intentional objects could be reformulated sententially, as declarative categorical propositions prefixed by an existential quantifier that predicates intentional objects like those of (1), (7), or (8) as properties, thus:

> (9) *(∃x)(x* is _____ [to go to the store, the number 3, the situation in Africa, etc.])[11]

But first, this suggestion fails for the intentional objects of conative attitudes like intending, hoping, desiring, fearing, etc. For the reasons just explicated, to intend to go to the store is not semantically equivalent to intending that there be something that is (my) going to the store. For similar reasons, to desire a piece of pie is not the same as desiring that there be something that is a piece of pie; nor is hoping for good weather the same as hoping that there is something that is

good weather; nor is fearing the plague the same as fearing that there is something that is the plague; and so forth. By asserting the independent existence of the intentional object, existential reformulations like (9) misrepresent such objects as ontologically and psychologically independent of the agent whose intentional object it is.

Second, (9) is like any other sentential formulation in that it *may* simply fail to represent the facts of the agent's actual intentional attitude, even in the easier case of the cognitive attitudes of thinking, believing, perceiving, conceiving, etc. For example, it may be true that I believe in magic, though false that I believe that there is something that is magic, and true that I perceive a dagger before me, though false that I perceive that there is something that is a dagger before me. Similarly, it may be true that I am thinking of the situation in Africa, though false that I am thinking *that* the situation in Africa has some particular property; or true that I am thinking of the number 3, though false that I am thinking anything in particular *about* the number 3; and so on. These intentional objects are alike, in that they can have no truth value independent of the truth value of propositions that ascribe the corresponding intentional attitude to the agent. Call these *agent-dependent intentional objects*. The intimacy of the relation between the agent, her cognitive attitude, and the agent-dependent intentional object of that attitude is disregarded by any such sentential reformulation in the manner of (9).

This is a significant oversight. An agent's ego or self is constituted, in part, by the cognitive and conative attitudes that define his conscious mental life. If *all* of those attitudes can take only intentional objects the truth values of which are independent of the agent's attitude toward them – call these *agent-independent intentional objects*, then none of the agent-dependent intentional objects just considered can constitute part of his mental life, nor, therefore, his sense of self. Nor can any of the agent's dreams, fantasies, disconnected memories, or free associations qualify, unless they can be formulated as propositions.

But this flies in the face of the psychological facts. Those of our dreams, memories, ideas, fantasies, and free associations that are most difficult to express sententially are often most personal, self-revelatory, and intimately constitutive of our selves. Indeed, nonsentential intentional objects are psychologically primary. We learn the singular terms, predicates and phrases that refer to them long before we learn the syntactical rules of grammar that anchor them in objective reality. Childhood fantasy depends on their potential for free-floating, ungoverned and arbitrary interpermutability, which transgresses the constraints of reality that syntax imposes. To learn the rules of syntax is gradually to abandon the daytime experience of their arbitrary interpermutability, except at those liminal moments when the mind begins to relax its grip on external reality in preparation for sleep, and properties and particulars that the waking mind rigidly separates begin to meld, merge and recombine in ineffable variation. And the prelinguistic interpermutational quality of dreams defies one’s subsequent attempts to capture them in language, the components of which may be subject to the same kind of displacement and arbitrary permutation. Thus sentential propositions themselves, and conventional grammar more generally, are inherently inadequate and unsuited to represent these most basic manifestations of

the self. They are equally insensitive to the poetic and literary tropes in which those manifestations find creative expression.

Some such mental contents are difficult to express sententially because of what we privately take to be their social unacceptability. But some are difficult to express sententially because they are simply not sentential propositions; and both poetry and ordinary speech, full of ellipses, ungrammatical breaks, and nonsyntactical strings and associations of words, reflects this. Moreover, it is often precisely *in virtue of* our inability to express certain thoughts sententially – and therefore in intersubjectively accessible form – that we think of them as exclusively our own. That they find a place in our internal lives but not a place in the external world of statable facts identifies them as such. Conversely, to express sententially intimate feelings or perceptions one shares with another is often to destroy their intimacy, and their status as private, personal, and shared; it may be to stifle them altogether, as Commander Data does his first girlfriend's desire that he kiss her on the neck by stating,

(10) I infer that you want me to kiss you on the neck.

– a wet blanket if there ever was one.

It is sometimes thought that it is the *verbal* expression of such feelings that sullies them. But not all verbalizations have this effect: poetry may not, song may not, disconnected murmurings or unfinished sentences may not. It is not the verbal expression of such feelings that is the culprit, but rather their agent-independent *sentential* expression. To formulate nonsentential intentional objects, and subject-predicate combinations of such objects sententially is, as Kant argued, to objectify and transform them, and to do this is to detach one's self from them. Kant rightly disputes Descartes' *cogito* on the grounds that I cannot infer the existence of my self from the activity of thinking with which it is identical: inferential relations can obtain only among suitably objectified sentential judgments, not between two mutually identical preconditions for making them [1C, B 422 – B 423fn].

Hence I agree with Kant that the copula "is" of declarative categorical judgments "is employed to distinguish the objective unity of given representations from the subjective," and that it is "not merely to state that the two representations [connected by the "is"] have always been conjoined in my perception, however often that perception be repeated; [but that] they are combined *in the object*, no matter what the state of the subject may be." [1C, B 142] I also agree with Kant's general argument, pressed strongly in the A as well as the B Deduction, that one cannot be a self without the ability to frame at least some of one's intentional attitudes in the form of *declarative* categorical judgments (see below, Section 4.3). But even this does not imply that the intentional attitudes *constitutive* of one's self consists solely, or even primarily, in attitudes towards intentional objects formulable in such sentential terms. If they did, Kant's synthetic function would have nothing to do.

To learn to objectify and transform nonsentential intentional objects into sentential form is part of the process by which we first come to recognize reality as independent of and external to our selves.[12] Freud thought that all such agent-dependent nonsentential intentional objects

were in some way constitutive of an agent's self; thus the importance to psychoanalysis of free association, slips of the tongue, and so on. I make only the weaker claim that *some* such objects have this function. These are the ones that most strongly resist public scrutiny in an impersonal idiom. Excluding these by definition or fiat from the scope of intentionality leaves us with an unnecessarily impoverished representation of an agent's ordinary mental life.

If nonsentential intentional objects that are psychologically fundamental to an agent's selfhood also may enter into the construction of sentential intentional objects, then they are among the constituents of sentential propositions agents can conceive *whether or not these propositions themselves contain intentional operators*. This is my second, converse proposal:

> *Proposal 2:* Anything that may function as a nonsentential intentional object may occupy the subject or predicate position in a sentential proposition that contains no intentional operator.

So, for example, "the situation in Africa" can function as a constituent in

> (11) The situation in Africa is intolerable

as well as it can in (8); "the number 3" can function as a constituent in

> (12) The number 3 has religious significance in many cultures

as well as it can in (7). But of course as Brandom and others have shown, sentential propositions themselves may function as constituents in more complex sentential propositions, whether the latter express intentional attitudes or not.

2.3. Intentionality and Subsentential Consistency

Following standard usage, I shall refer to subject and predicate constituents of propositions, both sentential and nonsentential, as *subsentential constituents*. Thus subsentential constituents are expressed by what Quine refers to[13] as *terms*. It will become evident that the main points I make here can be extended to cover the more complex subsentential constituents expressed by what he later redefines as *predicates*.

Further, I shall say that we have *concepts* of what both sentential propositions and their subsentential constituents correspond to in the world: complex states of affairs, and events and objects respectively, and properties of these; henceforth I refer to all of these collectively as "things". Basically, my notion of a concept follows Kant's account of the hierarchical relation between object, appearance, and empirical concept in the judgment, "All bodies are divisible," at 1C, A 68/B 93 – A 69/B 94 (also see 1C, A 109 and the discussion of 1C, B 104 in Section 1.3 above), which runs roughly as follows:

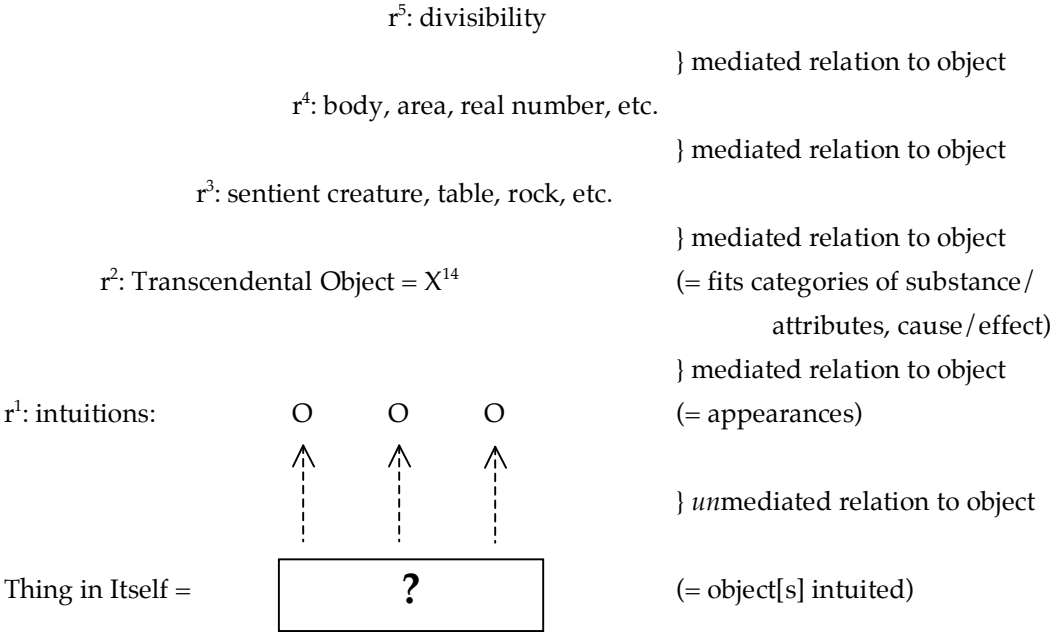


Figure 2. Kant's Conceptual Hierarchy

Earlier we saw that through intuition, according to Kant, we stand in direct and unmediated relation to unknowable states of affairs that are independent of the self, and make sense of the r^1-level representations we receive from them through the synthetic process of conceptual understanding. This process systematically combines r^2-level representations in a rule-governed way, such as to form concepts by which we identify these states of affairs as objects, i.e. as independent of ourselves as subject; and as having certain further, r^3- and higher-level empirical attributes. By thus conceptualizing them as objects, we trade unmediated relation to them for conceptual recognition of them. Because rule-governed synthesis is a precondition for recognizing anything as an object of experience at all, for Kant, no representation can enter empirical consciousness save as conceptually mediated, as Brandom rightly observes. Thus we necessarily conceptualize any object, including intentional objects, and in particular subsentential constituents toward which we take intentional attitudes.

We take intentional attitudes *toward* subsentential constituents; we have concepts *of* that which the resulting nonsentential intentional object represents. So, for example, we have concepts both

(i) that the number 3 has religious significance in many cultures [or, alternately: of the number 3 as having etc.]

and

(ii) of the situation in Africa.

In the following sections, I am concerned mostly with concepts, i.e. what Armstrong calls the "furniture of the mind,"[15] rather than with either the properties of the things we conceive, or the subsentential constituents that correspond to them. However, it will be convenient to approach the taxonomy of concepts through that of things, properties, and subsentential constituents themselves.

Do the requirements of theoretical rationality apply to subsentential constituents, whether or not those constituents are themselves sentential propositions? Consider sentential propositions such as the following:

(13) I intend to go to the store and not go to the store

It is tempting to think that we can explain what is wrong with (13) by analyzing it as a conjunction of two mutually contradictory propositions, thus:

(13a) I intend to go to the store and I intend not to go to the store.

But (13a) does not imply (13): That I have *two contradictory intentions* does not imply that I *intend a self-contradictory object*. Hence (13) and (13a) cannot be equivalent. Alternately, we might try giving (13) the form

(13b) I both intend that I go to the store and that I do not go to the store.

But (13b) as it stands is ambiguous. If "both" modifies "intend," then (13b) is really a compound proposition that includes two "intend that" locutions with contradictory objects, thus:

(13c) I both intend that I go to the store, and [intend] that I do not go to the store.

Hence (13c) again expresses two contradictory intentions, not one intention with a self-contradictory object. On the other hand, if "both" in (13b) modifies "that," then (13b) is really

(13d) I intend both that I go to the store and that I do not go to the store.

(13d) expresses *one intention with two mutually contradictory objects, not one intention with a single self-contradictory object*. For unlike (13), (13d) contains two intentional objects the truth value of each of which is independent of the agent's attitude toward them – i.e. that I go to the store and that I do not.

These are distinctions that HVTR is unsuited to make at the subsentential level. According to HVTR, the object of an intention is what follows the "intend that" locution, i.e. an atomic sentential proposition that comprises no further such propositions as constituents, or a compound sentential proposition that does ((13d) is of the latter kind). But in the weaker, ordinary sense, an object is merely a perceptually discriminable *thing*, i.e. anything that can be denoted by the subject term of a declarative categorical proposition. In this second, weaker sense of "object," a self-contradictory object is anything that can be denoted by a self-contradictory subject term, *whether or not that subject is itself a compound sentential proposition*. Thus, for example, we could rephrase (13) as

(13e) To go to the store and not go to the store is what I intend.

It is in cases like (13e), where the subject is *not* itself a compound sentential proposition that HVTR is not fine-grained enough to make the necessary distinctions. I have already argued that

subsentential constituents of propositions such as "to go to the store" are not uniformly semantically interchangeable with sentential propositions such as "I go to the store," because the latter are agent-independent, whereas the former are not. Where the self-contradictory subject is merely and irreducibly a conjunction of subsentential *constituents*, as in (13e), and not a conjunction of sentential *propositions*, HVTR has no conceptual resources for identifying the contradiction.

So we cannot explain what is wrong with believing propositions like (13), if the requirements of theoretical rationality apply only to the relations among sentential propositions we believe. For this is to say that the requirements of logical consistency apply only to those relations, and not to the relations among their subsentential constituents, whether we believe them or not. But this forecloses the obvious explanation of what is wrong with propositions like (13), namely that they contain *internal* logical inconsistencies in some as yet unexplicated sense. If we are to understand the centrality and inevitability of theoretical reason in the structure of the self, we need some way of explaining this natural reaction more systematically in terms of it. To this venture I now turn.

3. Rational Intelligibility and the Holistic Regress

The notion of the holistic regress, and the theoretically rational requirements of horizontal and vertical consistency introduced in Section 4 below draw heavily on Kant's conception of theoretical reason as developed in the Dialectic of *The Critique of Pure Reason*. See especially 1C, A 299/B 355 – A 308/B 364, A 322, 330-2, 337, B 378-9, 383, 387-88, 437, A 643/B 671 – A 669/B 697, "The Regulative Employment of the Ideas of Pure Reason"; compare 1C, B 93-4, 105-6 on judgments as functions for unifying our representations.[16] In what follows I do not claim to interpret Kant, but merely to develop and streamline some ideas that can be found in Kant's writings. Nevertheless, I try to navigate between the Scylla of technical issues in the philosophy of language and the Charybdis of Kant exegesis. My frequent references to Kant are thus intended to provide historical and motivational context for these ideas, not to represent them as what Kant actually meant (nor even, necessarily, what he should have meant).

I shall say that an event, object, or state of affairs (henceforth a "thing") is *rationally intelligible* to us if we recognize it as an instance of some concept. This definition of rational intelligibility draws on Kant's analysis of theoretical reason as inherently subsumptive and as similar in operation to the synthetic function of the categories of the understanding. But I also argue below that rational intelligibility implies logical consistency, hence theoretical rationality in a much weaker and more widely acceptable sense.

To *recognize* something is to perceive it as familiar, i.e. as the same as or similar to something you've perceived before. If something is in no respect like anything you've perceived before, then you cannot identify it at all. Does this imply that everything is rationally intelligible to us, since we recognize every thing as instantiating the concept of a thing? No, because the antecedent is false. From the fact that each thing does instantiate this concept, it does not follow

that we invariably recognize this. In Chapter VII, below, I examine some of the ways in which our theories about the world may thwart our recognition of the blindingly obvious.

As I use it here, the notion of recognition is a technical one, appropriated from Kant's account of concept-formation and -application. Briefly, Kant's idea is that we can identify something only if we have a concept of it; and can have a concept of it only if we can reproduce representations of it repeatedly in memory from moment to moment, and literally, *re-cognize* it at any given moment as the same as that which we cognized earlier, with respect to some property under the concept of which we subsume it. To do this is to conceive it as unified through time and so as an intentional object, with respect to whatever the particular properties by which we identify it.[17] To make something rationally intelligible, then, is to make sense of it as a discrete and unified thing; i.e. to conceive of it as independent of oneself as conceiving subject, by identifying it conceptually and thereby distinguishing it from oneself. In what follows I suggest the extent to which the requirements of theoretical reason must be satisfied in order for us to be able to do this.

Given some thing *t*, what must be true of us in order for us to have a concept of the kind of thing *t* is? Minimally, we must distinguish *t* from other kinds of things, not-*t*, which it is not. To do this we must recognize *t* as having at least one property, *P*, that things like *t* have, e.g. three-dimensionality, and that those other kinds of things lack. In order to recognize *t* as having *P*, we must have a concept of *P* and recognize *t* as an instance of it; or we must be able to acquire a concept of *P*, through experience or explanation, such that we would then recognize *t* as an instance of it. But in order to have or acquire a concept of *P*, we must be able to distinguish *P* from other kinds of properties, not-*P*, which it is not. To do this we must recognize *P* as having at least one higher-order property, $P1$, that properties like *P* have, e.g. length, and that those other kinds of properties lack.

I shall say that a property has a *higher order of comprehensiveness* than anything of which, as a matter of conceptual necessity, it must be predicated; and that our concept of that property has a higher order of comprehensiveness than anything that similarly must instantiate it. A property has a *lower order of comprehensiveness* than any property that, as a matter of conceptual necessity, must be predicated of it; and correspondingly, our concept of that property has a lower order of comprehensiveness than any concept it similarly must instantiate (I shall say more about conceptual necessity in Sections 4.1 and 5 below).

So, for example, having length has a higher order of comprehensiveness than having three-dimensionality. To recognize *P* as having at least one higher-order property $P1$, that properties like *P* have, e.g. length, and that other kinds of properties lack, we must have a concept of $P1$ and recognize *P* as an instance of it; or we must be able to acquire a concept of $P1$, through experience or explanation, such that we would then recognize *P* as an instance of $P1$. But in order to have or acquire a concept of $P1$, we must be able to distinguish $P1$ from other kinds of properties, not-$P1$, which it is not. To do this we must be able to recognize $P1$ as having

at least one higher-order property, $P2$, that properties like $P1$ have, e.g. being spatiotemporal, and that other kinds of properties lack. And so on.

Call this the *holistic regress.* The holistic regress is *holistic* because it implies that nothing can be rationally intelligible to us in isolation from things to which we recognize it as similar and other things from which we recognize it as differentiated. And it is a *regress* because it implies that in order for us to have a concept of the kind of thing some thing or property is, we must have or be able to acquire a whole host of further concepts of the higher-order kinds of thing that kind of thing itself is. For example, if we recognize something as three-dimensional, we also must be able to recognize it as having length, and moreover as spatiotemporal. If we are not able to recognize it as having these higher-order properties, we cannot recognize it as having the lower-order ones, either. This may not seem obvious, so I shall say more about it shortly. This account of the holistic regress implies that even if we were to encounter something we recognized as unlike anything else in the world, we could not understand in what respect it was unique until we'd encountered other things that, in sharing the property that made it unique, destroyed its uniqueness.

The Kantian holism I describe here stipulates relationships of contingent interdependence among the concepts an agent has at a particular moment. This kind of holism is different from the language holism of HVTR, which stipulates an inferential relationship among all constituents and sentences of a language as an interconnected matrix, such that any scheme that lacks that inferential interconnection must be atomistic rather than holistic. For example, Brandom endorses a Sellarsian brand of inferential concept holism that links having concepts with giving and having reasons that can justify beliefs and claims (MIE 89-90). But he thinks concept holism is independent of representationalism:

> [T]here is prima facie no reason why the fact that some object or property is represented by one simple idea, term, or predicate should be relevant to what is represented by others. Representational relations between nonintentional objects or properties and the intentional representings of them might be treated (as the empiricists in fact treat them) as separate building blocks that, when properly put together, determine what inferences are good in the sense of preserving accuracy of representation. Serving this role seems compatible with these presentational relations being quite independent of one another. Knowing what one state or expression represents need convey no information at all about what anything else might represent. (MIE 90)

However, if my argument above is valid, Brandom's view depends on a misrepresentation (so to speak) of what a concept is. Predicates are not the kind of thing that could hold of only one singular term, and my concept of it could not apply to only one instance of the thing that singular term denotes. The interpretation of concepts as representational does not reduce them to "separate building blocks that, when properly put together, determine what inferences are good in the sense of preserving accuracy of representation," because concepts

represent classes of objects that bear the relevant property and thereby distinguish themselves from others that do not.

The holistic regress has certain implications for the concepts with which we make the world and ourselves rationally intelligible to ourselves. First consider the *holism* of the holistic regress, i.e. the implication of it that we cannot recognize something as being of a certain kind, unless by comparison with other things to which it is similar, and by contrast with other things from which it is distinct, relative to certain properties. Clearly, such comparisons and contrasts imply satisfaction of the law of noncontradiction, i.e. that we cannot conceive a thing or property simultaneously as what it is and what it is not. Here what satisfies the law of noncontradiction is not the relation as we conceive it between things *and* their higher-order properties. So this requirement cannot be expressed by the relation between a predicate letter and the objects that fix its extension, thus:

(14) $(x)\sim(Fx \,.\, \sim Fx)$

What is required to satisfy the law of noncontradiction here are rather our concepts of the objects assigned to individual variables, i.e. our concepts of things and properties themselves. For this reason, I introduce here a few basic elements of what I shall call a *variable term calculus*, and develop this model at greater length in the following chapter.

Not just sentential propositions, but any rationally intelligible thing t assigned to an individual variable a must satisfy the following requirement:

(15) $\sim(a.\sim a);$

we must conceive it as self-identical, i.e. nonself-contradictory. So, for example, Quine's schematized axioms of identity

(I) $Fx.\ x=y.\ \rightarrow Fy$

(II) x=x

might be transformed into schematized axioms of nonself-contradiction, thus:

(I') $Fx.\ \sim(x.\sim y).\ \rightarrow Fy$

(II') $\sim(x.\sim x)$

One result of substitution of (I') would be, along Quinean lines,

(a) $z=x.\ \sim(x.\sim y).\ \rightarrow z=y$

from which would follow

(b) $\sim(z.\sim x).\sim(x.\sim y).\ \rightarrow \sim(z.\sim y),$

which we might call the law of transitivity of nonself-contradiction. The requirement of nonself-contradiction among terms and variables could function in proofs, as does the identity sign, either as an inert predicate letter or truth functionally with the insertion of an axiom of nonself-contradiction into the antecedent of the conditional. The holistic regress implies that we can recognize things and properties as nonself-contradictory only if we can identify them in terms of higher-order properties that are themselves nonself-contradictory.

4. Horizontal and Vertical Consistency

4.1. Horizontal Consistency

Next consider the sum total of things and properties that are simultaneously rationally intelligible to an agent at a particular moment, and the higher-order properties that make them so to her. Call the set *S* of concepts $c_1, c_2, c_3, \ldots c_n$ an agent has of these things and properties *the agent's perspective*. The relation between this limited set of concepts and the agent is something like the relation, according to Kant, between the concepts, both empirical and a priori, jointly necessary and sufficient for experience, and the "transcendental subject" whose concepts they are [1C, A 58/B 83 – A 62/B 87, A 127-8, B 165, 190-7, A 159, and especially A 651/B 679].[18] *S* includes concepts of properties of the external world, like length, as well as of the agent's own states, like desiring O or believing P or being in pain.

To say that *S* comprises an agent's perspective and not merely that of a static subject, abstractly conceived, implies that the agent's perspective changes over time, and with changes in her state, character, surroundings, and history. It evolves both progressively and regressively as the agent evolves over time, and may contain mostly[19] different members at one moment from those it contains at another. *S* as it is defined here comprises only those concepts by which the agent *actually does* make things rationally intelligible at a particular moment, not the ones by which she *could have* made them so, nor any other concepts she has at her cognitive disposal. To this extent the concepts that constitute an agent's perspective *S* at a particular moment in time are occurrent, but need not be linguistically explicit or manifest in overt behavior.

Agents' perspectives differ with respect to the things and properties of which they have concepts (this is one reason why people sometimes find each other incomprehensible), and differ also with respect to the scopes of instantiation of those concepts (this is one reason why people who share the same assumptions and vocabulary often disagree with or misunderstand each other), and so with respect to the conceptual necessity of their instances. For example, most of us would probably agree that a three-dimensional thing instantiates, as a matter of conceptual necessity, the concept of a thing's having length; but would show less consensus that going to the store instantiates, as a matter of conceptual necessity, the concept of a tedious errand. Those of us who go to the store infrequently may think of it instead as an entertaining diversion; others may not think of going to the store as instantiating, *as a matter of conceptual necessity*, any concept, not even that of an action. Because we each may have different perspectives on such matters, the definitions just offered of higher and lower orders of comprehensiveness must be relativized to an agent's perspective.

Whatever the sum total of concepts that constitute my perspective at a particular moment, the holistic regress implies that the law of noncontradiction must be satisfied simultaneously by all of them. Otherwise there would be some thing or property I could neither identify with nor differentiate from anything else. In that case I could neither identify any of those other things with it, nor differentiate any of those other things from it. And then I could make none of them rationally intelligible. This is to say that I must conceive all the things and

properties that are simultaneously rationally intelligible to me as logically consistent with one another; i.e. that

(A) *S* observes the law of noncontradiction, in that the members of *S* are internally and mutually consistent in their application.

(A) makes the requirement of nonself-contradiction stated in (15) a special case of the familiar law of noncontradiction more generally. (A) says that we can understand particular things or states of affairs only if the concepts by which we recognize them are neither internally nor mutually contradictory. In standard notation modified as suggested above, (A) would run roughly as follows: For any agent's set *S* of concepts of things and properties $c_1, c_2, c_3, \ldots c_n$, and rationally intelligible things or properties $t_1, t_2, \ldots t_n$ assigned to individual variables $a_1, \ldots a_n, b_1, \ldots b_n, \ldots$,

(HC) $(\sim\exists x)(x.\sim x)$,

i.e. we must conceive any such c_i as self-identical, i.e. nonself-contradictory.

Call this the requirement of *horizontal consistency*. For now, some readers may wish to read the expression enclosed in the second set of parentheses in (HC) as predicating "$.\sim x$" of x. But I discuss (HC)'s notational peculiarities at greater length in Chapter III.5, 7 and 9, below.

4.2. Vertical Consistency

Next consider the *regressiveness* of the holistic regress, i.e. its implication that we cannot have concepts of the kind of thing some thing or property is, without being able to invoke further concepts of the higher-order properties that in turn identify that kind. This means that if I recognize some thing or property as a certain kind of thing, I also must be able to conceive it as of the same higher-order kind as is the kind I originally recognized it to be. So, for example, if I recognize something as a three-dimensional thing, I also must be able to conceive it as a thing of a certain length; if I recognize going to the store as a tedious errand, I also must be able to recognize it as nothing extraordinary. Otherwise it would be possible for me to recognize something *t* as having a certain property *P*, and also as having the negation of some further property $P1$ that *P* implies. In that case I would not have succeeded in making *t* rationally intelligible in terms of *P* in the first place. More generally, I must conceive the higher-order properties by which I recognize something as logically entailed, as a matter of conceptual necessity, by the relevant lower-order ones. This is to say that

(B) any particular c_i in *S* is either

(1) an instantiation of some other c_j in *S*; or

(2) instantiated by some other c_k in *S*;

i.e. *S* is minimally coherent;

(C) for any cognitively available particular thing *t*, there is a c_j in *S* that *t* instantiates, i.e. *S* is complete.

(B) says that the concepts that constitute my perspective *S* are minimally coherent with one another, in that each particular thing identified by them satisfies the subject-predicate

relationship with respect to at least one other of them. (C) says that *S* is complete, in that any particular thing itself of which I am conscious instantiates at least one of them. Call this the requirement of *vertical consistency*. In standard notation, the requirement of vertical consistency would run roughly as follows: Given an individual variable *a* to which *t* is assigned, and terms *F* and *G* with the extensions *P* and P^1 respectively,

(VC) $Fa \rightarrow [(x)(Fx \rightarrow Gx) \rightarrow Ga]$

It is important not to confuse the requirement of vertical consistency with a claim about the transitivity of predication generally: Not every property is of a higher or lower order than every other property. The claim is not, for example, that if the pencil is red and red is fashionable this year, that the pencil is therefore fashionable this year. For not all red things are fashionable this year (e.g. firetrucks, blood). Rather, the requirement of vertical consistency is a transitivity claim about the relation between lower- and higher-order properties, i.e. those that satisfy (VC). It implies simply that the relations between our concepts of the lower-order properties of a thing and of the relevant higher-order ones are transitive: If the pencil is three-dimensional, and three-dimensional things have length, then the pencil has a certain length. So if I recognize the pencil as three-dimensional, and three-dimensional things as having length, then I recognize the pencil as having length.

Also notice that vertical consistency does not require that I be able to recognize something as having *all* the higher-order properties that in fact apply to it; just that the ones by which I do recognize it be implied by the relevant lower-order ones by which I recognize it. In Section 6 below, I argue that there must be at least one such higher-order property in order for me to recognize it as anything at all. Nor does vertical consistency require that I be able to recognize the relations that obtain between a thing, its properties, and the further properties that they have but that the thing does not (e.g. such that the pencil is not a primary color although red is). Because the requirement of vertical consistency applies only to the relations among properties that satisfy (VC), there may be "floating hierarchies" which are unconnected to others within an agent's perspective. However, I argue in Section 6 that all of them must be related as (VC) describes to the highest-order property that defines this perspective as an *agent's* perspective.

The requirements of horizontal and vertical consistency systematize and unify the set *S* of concepts constitutive of an agent's perspective at a particular moment. (HC) and (VC) ensure that, whatever the concepts constitutive of *S* at that moment, they will be mutually rationally intelligible. However, (HC) and (VC) do not ensure, either separately or conjointly, the persistence through time of any such ordinary concept. It is consistent with the satisfaction of (HC) and (VC) at each moment in time that the concepts constitutive of *S* at t_1 are almost entirely disjoint from those constitutive of *S* at t_n. I qualify this claim in Section 6 and Chapter III, below; but it holds for most ordinary concepts. Envision, for example, the effect on *S* of constant and instantaneous transmission of global information, simultaneously with sudden and pervasive paradigm shift in the natural sciences. Practically everything could change very quickly, and very

traumatically, with correspondingly traumatic consequences for an agent's perspective. Less traumatic changes in an agent's perspective are to be expected in the normal process of growth and evolution of character and circumstance.[20]

4.3. Kant on Horizontal and Vertical Consistency

(VC)'s similarity to modus ponens is not accidental. Versions of both are to be found in Kant. But by contrast with my formulation of (VC), Kant attempts to insure satisfaction of the requirement of vertical consistency by proposing his Table of Categories as comprising a priori necessary conditions of any *kind* of judgment we might make. He tells us repeatedly that if a perception does not conform to the fundamental categories of thought that ensure the unity and coherence of the self, they cannot be part of our experience at all [1C, A 112, 122, and B 132, 134]. This thesis may be viewed as the resolution of a *Gedankenexperiment* he earlier conducts at 1C, A 89-91, in which he entertains the possiblity of unsynthesized appearance.[21] In any case, his ultimate commitment to this thesis is clear. Kant describes these fundamental categories as "*a priori* transcendental concepts of understanding," by which he means innate rules of cognitive organization that any coherent, conscious experience must presuppose.

The table of transcendental categories Kant offers in the Metaphysical Deduction is drawn largely from Aristotle, with his own considerable additional tinkering. The categories include substance, totality, reality, possibility, causality, and community, to name just a few. But some commentators[22] have rightly concluded that the most significant candidate for this elevated cognitive status is the subject-predicate relation in logic, from which Kant derives the relational category of substance and property in the Table of Categories (Kant regards this as the result of fleshing out the subject-predicate relation or judgment form with transcendental content, i.e. the sensory data our experience presupposes rather than the sensations we perceive as a result of it. [1C, A 70/B 95-A 79/B 105]). The idea, then, would be that organizing sensory data in terms of this relation is a necessary condition of experience. On this view, if we do not experience something in a way that enables us to make sense of it by identifying properties of it, we cannot consciously experience that thing at all.

This neo-Kantian revision has the merit of plausibility over the archaic list of categories Kant originally furnished, for it is simpler and more noncommittal on the issue of to what extent our cognitive capacities are hard-wired, and what their content must be. It does not seem too controversial to suppose that any viable system of concepts should enable its user to identify states of affairs by their properties, since concepts just are of corresponding properties, and to ascribe a property to an object just is to subsume that object under the corresponding concept. So any system of concepts should enable its user to ascribe to objects those properties of which she has concepts.

My proposed requirements of horizontal and vertical consistency are a further extension of this neo-Kantian revision. They are weak enough that they may even be defensible in the face of anthropological evidence that languages considerably remote from Indo-European ones evince

a cognitive structuring to the user's experience that is so different from our own as to be almost unintelligible to us. It would be an argument in favor of (HC) and (VC) if it could be shown that the subject-predicate relation held regardless of the other ways in which culturally specific conceptual organizations of experience differed among themselves.[23] (HC) and (VC) imply that if we do not experience something in such a way as to allow us to make sense of it in terms of a set of coherent concepts that structure our experience at a particular moment, *whatever those concepts are*, we cannot consciously experience that thing at all. On this thesis the innate capacity would consist in a disposition to structure experience conceptually as such, but not necessarily to do so in accordance with any particular list of concepts,[24] provided that the particular, culturally specific set S of concepts c_1, c_2, c_3 … c_n that did so satisfied (A) – (C), i.e. (HC) and (VC).

These two requirements, of horizontal and vertical consistency, illuminate further the sense in which nonsentential intentional objects are psychologically fundamental in the structure of the self. In Section 2.2 I claimed that nonsentential intentional objects do not necessarily imply the agent-independence of that which they represent from the agent whose intentional objects they are. The holistic regress implies that in order for the question of a thing's agent-independence to arise, one must first have made that thing – be it event, particular, state of affairs, or mistaken perception of any of these – rationally intelligible to oneself. And one can do that only by conceiving it in a way that satisfies the requirements of horizontal and vertical consistency. A thing must be rationally intelligible to us *before* we can formulate declarative propositional beliefs about it; a close look at Kant's account of concept-formation and -application, particularly in the A Deduction, and his rather obscurely argued claims in the Dialectic as to the relation between intuition, understanding, and reason, might show this to be Kant's thesis as well.[25] It would therefore be unilluminating to explain the rational intelligibility of a thing to an agent by imputing sentential beliefs to that agent.

4.4. The Interdependence of Horizontal and Vertical Consistency

It may not seem necessary to satisfy both horizontal and vertical consistency. It may seem that I could recognize a thing as having some lower-order property, as similar to other things that have that property and different from other things that lack it (i.e. requirement (A)), without that property itself being rationally intelligible to me in terms of some higher-order property it has at a given moment (i.e. requirements (B) and (C)). In that case, the requirement of horizontal consistency would be satisfied, although that of vertical consistency did not apply. Thus, for example, in the early stages of concept-formation, an infant may be able to recognize certain things as three-dimensional, without being able to recognize three-dimensional things as spatiotemporal. At the same time, I could not have concepts of the lower- and higher-order properties by which I recognize something, without simultaneously having other concepts of what they are not. So in theory, it may seem, my concepts of the things that are rationally intelligible to me at a particular moment may be horizontally consistent without being vertically consistent, but cannot be vertically consistent without being horizontally consistent.

However, it is not possible for the concepts that constitute my perspective to be horizontally consistent without being vertically consistent. Suppose, for example, that we were to be confronted with some particular thing such that the concepts it instantiates satisfied (A) but violated (B) and (C), i.e. such that we could invoke a concept in identifying it consistently with the application of our other concepts; but that that concept itself bore no instantiation-relation to others in the set (i.e. aside from that trivial one of being a concept in the set). In this case, that which we invoked as a "concept" would in fact not be one at all, since the corresponding predicate would by definition denote only the single state of affairs it had been invoked to identify. Since there would be no further concepts in terms of which we might understand the meaning of that denoting term, it could not enter into any analytic truths. In short, this would be like cooking up a special noise to denote only one state of affairs on the single occasion of its occurrence; precisely thus can the prelinguistic noises of infants be interpreted. In such a case the enterprises of denotation and meaning themselves would fail.

Similarly, it is not possible for the concepts that constitute my perspective to be vertically consistent without being horizontally consistent. Imagine, for instance, what it would be like to be confronted by a particular thing such that its concept satisfied (B) and (C) but not (A), i.e. such that it enabled us to identify its properties in terms of concepts in the set, but the application of those concepts themselves was internally or mutually inconsistent. In that event, it would be possible to violate (VC), i.e. to ascribe to the thing the conjunction of some predicate *F* and some other one, *G*, that implied the negation of *F*. Again the enterprise of identification itself would fail. If we were finally to fail to identify the thing or state of affairs in question as having a consistent set of properties, we would fail to identify it altogether. And then it could not be part of our conscious experience.

For example, a friend of mine – let's call her Joan – related the following true story. One night while she was lying in her bed in Cambridge, Massachusetts, reading, her bed tipped sharply upwards. As Joan describes this happening, she "immediately forgot that it happened." She did not "remember" that it had happened until a few days later, when she heard on the weather report that, at that very moment, New England had experienced its first major earthquake in decades. Joan accounted for her "amnesia" by saying that because she had had no possible explanation for her bed tipping, as far as she was concerned the event had not happened.

I would suggest that her account was almost right, but too strong. First, it is not that she was momentarily conscious of her bed tipping and then forgot it until she found the appropriate explanation. After all, how could one simply forget such a momentously anomalous event, merely for lack of an explanation of it, when one would have thought it would be precisely its cryptic and inexplicable character that would fix it in one's mind? My proposed account is different. Rather than having forgotten her bed tipping upward, Joan did not consciously experience that event in the first place, even though it happened to her. Second, it was not an *explanation* she needed in order to register that event as an object of her experience. Rather, she

merely needed a relevant higher-order *concept* that enabled her intelligibly to identify it as having happened to her earlier. Keep in mind that among the concepts that constitute an agent's perspective are concepts of properties of things. So if you do not have any higher-order concepts under which to subsume the event, you cannot even ascribe properties to it. It often happens that we do not register certain events in consciousness until long after the fact, when some relevant concept or conversation first calls them to mind and enables us to identify them.

Now Joan clearly had the concept of her bed tipping in her arsenal of *possible* concepts. Why wasn't that sufficient to enable her to maintain the event in memory? Why couldn't she simply have predicated of her bed that it was tipping? And why wouldn't that have been sufficient for her to have made it rationally intelligible to herself? My answer would be that the event in question violated the vertical consistency of her perspective: although she had the concept of the property of her bed tipping, there was no relevant higher-order concept available under which she necessarily could subsume that one. There was simply no room for it within her conceptual scheme.

A similar explanation could be offered of more traumatic, conceptually anomalous events that may happen to an agent, such as war or childhood sexual abuse; as well as of normal early childhood amnesia. Freud explains our failure to remember the events of early childhood by the concept of repression. I suggest instead that we simply lacked the concepts by which to identify them. To the extent that we are lucky enough to learn the right ones now, we may "remember" – i.e. make rationally intelligible – those events, just as Joan did the tipping of her bed. The general phenomenon of remaining unconscious of things accessible to an impartial observer is commonly called *denial*. I discuss it at greater length in Chapter VII below. Denial functions to maintain vertical consistency within an agent's perspective against the threat of external cognitive anomaly.

If such cases characterized all of our encounters with the world, we would have no experiences of it at all, and therefore no unified sense of self either. These are the sorts of failures Kant has in mind when he avers, in the A Deduction, that

> without [the synthetic unity of appearances according to concepts], which has its a priori rule, and subjects the appearances to itself, no thoroughgoing and universal, therefore necessary unity of consciousness in the manifold of perceptions is to be found. These [perceptions] then would not belong to any experience, therefore would be without an object, and nothing but a blind play of representations, that is, less even than a dream. [1C, A 112]

In this passage Kant sketches – for the first time, to my knowledge – the idea of an unconscious, in which extant perceptions are not rationally structured by the demands of external reality. Kant is saying that if we do not organize cognitively the data of our senses according to consistent and coherent rules, we cannot be rationally unified subjects. "For otherwise," he adds in the B Deduction, "I would have as many-colored and diverse a self as I have representations of which I am conscious." [1C, B 134] I would, that is, lack a sense of myself as the subject in whose

consciousness those representations occur. This is the sense in which, for Kant, the cognitive organization of experience according to consistent and coherent concepts is a necessary condition of being a rationally unified subject. In Section 6 I argue that an agent whose perspective fails to satisfy the requirements of horizontal and vertical consistency cannot exercise her agency at all. The observed behavior of infants would be consistent with this argument.

5. Intentionality, Consistency and Rational Intelligibility

In conjunction with the claims defended in Sections 2.1 – 2.3, the requirements of horizontal and vertical consistency enable us to say in somewhat greater detail what we instinctively find wrong with (13). If we accept the argument of Section 2.1, that some objects of intentional attitudes are nonsentential, then we can treat those attitudes straightforwardly as *properties* that may be ascribed both to the events and objects that constitute the complex states of affairs to which sentential propositions correspond, as well as to those complex states of affairs themselves. For example, to go to the store may have the property of my intending it, just as the situation in Africa may have the property of my thinking of it, or my doubting that P may have the property of my experiencing or desiring it.

Now *my going to the store* does not *necessarily* have the property of my intending it: I could conceivably intend just the opposite under those same circumstances. But going to the store is something I now intend only if it is something I now occurrently conceive; i.e. only if going to the store is the object of a concept that is part of my current perspective. But something is the object of a concept that is part of my current perspective only if it necessarily has the property of my occurrently conceiving it: I can conceive of no particular thing that lacks the property of my conceiving it – neither the situation in Africa, nor my doubting that P, nor going to the store, nor anything else; nor will I ever be able to do so. Everything I ever conceive necessarily will have the property of my conceiving it. So going to the store is something I now intend only if it has the property, as a matter of conceptual necessity, that I now conceive it.

But there is no intentional attitude that consists simply in my conceiving something, irrespective of how I conceive it (not: irrespective of what I conceive it to be). Rather, I conceive it *as a certain kind of intentional object*: of faith, or fear, or intent, or desire, or belief, or contemplation, or curiosity. That is, the intentional attitude I take toward the thing is contained in my concept of it. Like any intentional object, *my occurrent concept of* my going to the store contains, as a matter of conceptual necessity, the higher-order property of the intentional attitude I take toward my going to the store. This is not to imply that my occurrent concept of, for instance, the situation in Africa necessarily includes *the concept of* my deploring it. But it does necessarily include *my deploring it*. This substance-property relationship then may be expressed in declarative categorical propositions such as

(16) To go to the store is what I intend.

Hence intentional objects and our attitudes toward them are subject to the requirements of horizontal and vertical consistency.

Now review these requirements of rational intelligibility: I make something rationally intelligible by recognizing it as a certain kind of thing. According to (16), I recognize going to the store as an intention I have. But if we rephrase (13) as declarative categorical proposition

(13e) To go to the store and not go to the store is what I intend,

we find that we clearly cannot predicate anything of its subject, because that subject violates the requirement of horizontal consistency. The subject of (13e) describes an event that is both what it is and what it is not. And we already know that we can recognize no such event as rationally intelligible in the first place.

Here are some further examples of self-contradictory intentional objects that violate the requirement of horizontal consistency but are invisible to the propositional view:

(17) My strongest gustatory desire is for the martini and not the martini.

(18) Clive and not Clive is the best cyclist in town.

Now HVTR would no doubt respond to these further examples by attempting the same sort of sentential reduction as it has for (13), and I would respond by mounting against them the same sorts of objections as I already have. But one final consideration against the primacy of sentential reduction may furnish at least an intermission in the debate. This is the spectre of an infinite regress of such reductions; a regress far less benign than the holistic one. If even atomic subsentential constituent intentional objects, like those in (1), (7) or (8) can be reduced to sentential judgments, it is difficult in principle to see how we can ever accurately identify the mental states and cognitive processes necessary in order for us to learn to construct such judgments in the first place.[26] HVTR might retort that if I recognize something as a certain kind of thing, it is surely to make a judgment, e.g. "That is a football." But my point is that the recognition of the thing is a *necessary condition* of making the judgment, not *identical* to it. If I could not first ponder the application of the indexical concept of thatness, and envision to myself a football hurtling through the air, I could not learn to make the propositional judgment at all. And I submit that although the intentional object of the attitude expressed in the following sentential proposition

(19) I envision a football hurtling through the air

is perfectly intelligible to us, there is no sentential reduction of the constituent, "a football hurtling through the air," that makes it so.

So far I have argued that the requirements of horizontal and vertical consistency are implied by the holistic regress, and that the holistic regress, in turn, is implied by the requirement of rational intelligibility. The further implication of this argument is that if we are successfully to make coherent sense of things, even in the most minimal way, we must, in conceiving those things, satisfy the law of noncontradiction in the ways the requirements of horizontal and vertical consistency specify. This is the sense in which, I want to claim, the minimal consistency requirements of theoretical reason apply not just to sentential propositions, but also, and more fundamentally, to those concepts of their constituents that form both an agent's perspective, and so her self. But if the concepts that constitute an agent's perspective, whatever they are, must

satisfy the requirements of horizontal and vertical consistency in order that the world be minimally rationally intelligible to her, then whether an agent is theoretically rational or not cannot depend upon contingent factors, such as training or personality, that some normal human agents have and others lack. An agent who is not theoretically rational in the minimal sense to which the requirements of horizontal and vertical consistency commit us cannot make sense of the world at all.

Now it might be objected that I have made my point only by changing the subject; and that this minimal sense of "theoretically rational" is not the one we ordinarily have in mind when we ask whether or not an agent is theoretically rational, and in virtue of what characteristics he is or is not. But the requirements of horizontal and vertical consistency are in essence the same rationality requirements we ordinarily do have in mind when we ask these questions, namely the requirements of logical consistency. Since any sentential proposition itself can be embedded in another one as a constituent, the requirements of horizontal and vertical consistency can be applied as well to sentential propositions and strings of such propositions, to yield the familiar canons of theoretically rational inference to be found in any logic textbook: Sentential propositions that satisfy the requirement of horizontal consistency thereby satisfy the requirements of sentential logic, and sentential propositions that satisfy the requirement of vertical consistency thereby satisfy at least some of the (less controversial) requirements of quantificational logic. The requirements of horizontal and vertical consistency are therefore not qualitatively different from the familiar ones. My objective in spelling them out has been to frame these familiar canons in such a way as to call attention to their applicability, not just to complex premises, arguments, and theories, but also to the most basic concepts in terms of which we understand the world around us. The implication is that all normal human agents are theoretically rational to some degree.

6. The Self-Consciousness Property

Next I take up what may seem to be some obvious objections to the claims defended in the preceding sections. First, is there really a holistic regress in the concepts by which we make sense of things? Why could we not minimally understand a number of different things by recognizing each as having just one, or a few lower-order properties? Or, more plausibly, perhaps: Why can we not more fully understand many different things in the world, ultimately in terms of a few, very comprehensive categories – life, death, human nature, physical forces, say – that themselves cannot be made rationally intelligible in terms of any more comprehensive ones?

First we must keep in mind that the question is not about the higher-order, comprehensive properties that may *in fact* sort things in the world into natural kinds. Instead, it is about what conditions are necessary so that *we* can make these things rationally intelligible to ourselves. Kant's answer to this was that we are naturally disposed to the holistic regress by the nature of our theoretical reason itself, to ask repeatedly for increasingly comprehensive, unifying

principles by which to identify and explain things; to subsume them under higher-order, increasingly comprehensive concepts; and finally to "cap" the regress by subsuming them all under the highest-order concepts of God, freedom, and immortality [1C, A 299/B 356 – A 314/B 371, A 321/B 378 – A 328/B 385, A 330/B 387 – A 341/B 399]. My own embellishment on Kant's answer is to argue that he was right about the holistic regress, but wrong about the highest-order concepts to which it inevitably leads us.

Suppose I did sort my experiences into the higher-order concepts of life, death, human nature, and physical forces, without recognizing those things as instances of some yet higher-order concept. Recall first that one advantage of acknowledging nonsentential intentional objects was that intentional attitudes then could be conceived as properties of the things to which such intentional objects correspond. Now an intentional attitude is a property of the thing I have the intentional attitude toward, whether or not I am empirically self-aware of my own intentional attitudes. A concept can be part of my current perspective even though I am not empirically self-aware of it as such. For any such concept need only be occurrent. It need not be explicit. A concept constitutive of my perspective is *explicit* if I am empirically aware of it at that particular moment. But to be *occurrent* it need only be in use at that moment. A concept can be currently in use at a particular moment although I am not empirically aware of it at that moment. Hence a concept can be both occurrent and also *implicit* at a particular moment if it is in use but not a current object of empirical awareness. And more specifically, a concept of my own intentional attitude at a particular moment is both occurrent and implicit, if it is in use but not, at that moment, an object of my empirical self-awareness. Concepts of our own intentional attitudes often have this feature. So believing, desiring, intending, thinking, etc. also can be treated as properties of which agents have occurrent but implicit concepts. Thus, for example, I may have an occurrent and explicit concept of my desiring an entertaining diversion; or, if my desire is unconscious, an occurrent but implicit one. In either case, we can think of my desire as a property of the envisioned entertaining diversion.

Now if I could not conceive life, death, etc. as instances of some higher-order concept, then in particular, I could not conceive them as instances of *my experience*: I could not recognize each of the things I identified as life, death, etc., as having the further property of being an object of an experience I had. I shall refer to the property of being an object of an experience I have as *the self-consciousness property* of things I in fact experience; and henceforth reserve the term "self-aware" to denote the case of explicit, empirical and contingent awareness of one's intentional attitudes. This account of the self-consciousness property departs from Kant's uncertain and conflicting pronouncements on the status of the "I". Sometimes he seems to think it is a concept [1C, B 133, n., 134, 423, n., 428-30, A 341/B 399 – A 342/B 400, A 348, 400]; and sometimes not [1C, B 68, A 117, n., 382, B 423, n.]. I think not only that it is a concept, but (as we shall see) that it deserves the status of a *concept of reason*, as Kant characterizes that notion [1C, A 310/B 367 – A 311/B 368], in virtue of its "contain[ing] the form of each and every judgment of the

understanding and accompany[ing] all categories as their vehicle" [1C, A 348]. Kant comes close to acknowledging this at 1C, A 682/ B 710.

That an experience has the self-consciousness property does not (to repeat a caveat from this chapter's introduction) require its agent to engage in explicit empirical reflection on it. So it does not require perpetual empirical self-awareness of the sort that most of us have to work quite hard to maintain. The self-consciousness property of my experiences is occurrent, but may well be implicit rather than explicit most of the time: It requires only that I be capable of identifying the experience as mine, not that I in fact do so of every experience I have.

Among the objects of my experience are both things in the world and my own intentional states. By hypothesis, I recognize each of these as instances of life, death, etc. But if I could not then recognize each of the things I identified as instances of life, death, etc. as in turn having the self-consciousness property, I could not conceive of any of these things as objects of my experience. Of course this does not mean that they would not *be* objects of my experience; just that I could not conceive them as such.

But an agent who lacked the concept of the self-consciousness property even implicitly would lack the capacity to recognize herself as partially responsible for the character of those experiences – and so, finally, would lack a necessary condition for motivationally effective agency.[27] Consider what such an agent might be like. She might have concepts of properties that attach to the *de facto objects* of her experiences, i.e. to the events, objects, and states of affairs she experiences, for example, being human nature, or bright red, without thereby having the concept of herself as *subject* of them. In this case, she would regard a characteristically human or bright red object of experience impersonally as occurring, but not as occurring *to* anyone.

Alternately, she might have, in addition, concepts of properties that attach to herself as a *subject* of experience, i.e. to the way she experiences such events, objects, and states of affairs, for example, being surprised by something, or open-minded to something, or desiring something. She would have to regard such intentional states of surprise, or open-mindedness, or desirousness as *happening to her*. But she would not necessarily regard them as *her* states. Instead, she might feel involuntarily overtaken by surprise, or stripped of her opinionated defenses, or propelled by desire, in spite of her character dispositions and impulses. In this case, she would view these states as alien and invasive psychological forces that happen *to* her, but not, as it were, *from* her. Most of us experience this sense of powerlessness over and detachment from our own intentional states at some point. Many feel this way about sexual attraction, or compulsive gambling. Others may claim to feel this way about all experience; that is, they may take the naive realist view that the character of a particular experience they have is entirely dependent on the character of its objects, and not at all on that of its subject.

In order for an agent to regard his experiences of different things as objects of his experiences, he must be able to recognize such experiences not only as occurring in just that form exclusively to him, but *a fortiori* as doing so *in virtue of his nature*. That is, he must be capable of viewing such experiences as not only *affecting* him, but also as being partly *determined by* him.

Thus he needs to be able to recognize his experiences as the result of an active, reciprocal collaboration between their subject and their objects, and as having the particular character they do in virtue of that collaboration.

But even the naive realist just described must grant this much. For without an implicit recognition of one's collaboration in the character of one's experiences, one would lack a necessary condition of being motivated intentionally to alter those experiences, i.e. to act. By hypothesis, in lacking the concept of the self-consciousness property, such an agent would not necessarily lack higher-order concepts under which all lower-order ones might be integrated by the requirements of horizontal and vertical consistency. But without the highest-order concept of their being objects of *her* experiences, their rational intelligibility would not be recognized as depending in any way on *her* behavior or condition, nor as susceptible to any attempts of hers to preserve it. Hence although her perspective might, quite fortuitously, satisfy the requirements of horizontal and vertical consistency at a particular moment, she would be unable intentionally to mobilize the psychological resources – i.e. the acts of attention Kant maintains (1C, B 68-69, 140, 153-6, 157-8a) are essential – for sustaining it in that form from one moment to the next.

For example, she might interpret the experience of forgetting, rather, as a temporary lacuna in the objective history of events. She might interpret her experiences of inference or theory-building as a direct perception of nonmaterial processes. She might view her most intimate processes of thought and feeling as external conditions visited upon her over which she had no control. And she would experience actions as involuntary behavior, propelled by external teleological forces to which she was subject. Thus she would lack agency, not just in the ordinary sense of being incapable of gross physical action. She would lack it as well in the more pervasive sense, in which we ordinarily conceive ourselves actively to *do* things like think, feel, infer, and search our memories.[28]

Without the concept of oneself as having one's experiences, everything would be conceived as being done *to* one, and nothing *by* one. So we must have some such degree of self-consciousness in order to sustain not only some minimal degree of rational intelligibility, but our agency as well: Each thing, therefore, that is rationally intelligible to me at a given moment must, *as a matter of conceptual necessity*, instantiate the concept of an object of my experience. For without the concept of the self-consciousness property, our perspective on the world would not be an *agent's* perspective at all.

I should particularly like to press this conclusion on Humeans who stubbornly avow – despite my best efforts in Volume I – that all action is motivated by desires, to the satisfaction of which theoretical reason is merely instrumental. No matter how vacuously the concept of desire is construed, this cannot be right if a motivationally necessary condition of action of any kind is the implicit theoretical conception of oneself as having the desire, or aversion, or resolution, etc. in question. The representational analysis of desire I developed in Volume I, Chapter II.2.1 satisfies this necessary condition, whereas the unreconstructed Humean notion of desire does not.

Thus the self-consciousness property as just explicated satisfies (VC): If I recognize the pencil as three-dimensional, and three-dimensional things as objects of my experience, then I recognize the pencil as an object of my experience. Does this entail that I recognize Socrates' death as an object of my experience because I read about it? Yes: things I read about become *objects* of my experience, even if I do not *immediately experience* them. This is because as I use the concept, all objects of my experience are inherently theory-laden – to different degrees and often with different theories (see below).

7. Intelligibility and Transpersonal Integrity

That we must finally be able to invoke the concept of the self-consciousness property in order to make things rationally intelligible answers simultaneously three further questions about the holistic regress. First, it answers the following objection. If I cannot make anything rationally intelligible without recognizing it as an instance of some higher-order concept, then unless my perspective includes an infinite number of higher-order concepts, it is hard to see how I can make rationally intelligible anything at all. And since my perspective at any particular moment clearly does *not* include an infinite number of higher-order concepts, either I do not make things rationally intelligible, or else the profferred account of how I do so must be wrong.

Now we can easily imagine some higher-order concept we might invoke, in turn, to make rationally intelligible the concept of the self-consciousness property. The concept of an event in the world, or of a sentient occurrence, for example, are both instantiated by objects of my experience. So it might seem that the concept of the self-consciousness property supplies no proper termination to the holistic regress. But in order for the concept of an event in the world itself to be rationally intelligible to me, I must be able to recognize things as events; and to do so requires, minimally, that they, too, be objects of my experience. So in order for the concept of an event in the world to be rationally intelligible to me, it must instantiate the concept of an object of my experience, and not the other way around. And similarly with all such concepts. So the concept of the self-consciousness property is of a higher order than any other in an agent's perspective.

The highest-order concept of the self-consciousness property terminates the holistic regress by rendering it innocuous. For this regress is then just the familiar regress of self-consciousness: For any object of my experience E, there is an object E^1 of my experience of E, and an object E^2 of my experience of E^1 of E, and so forth. That the holistic regress resolves into the infinite regress of self-consciousness just means that ultimately, we must be able to make everything, including objects of our experience, and objects of objects of our experience, and so on, rationally intelligible to ourselves as objects of our experience – i.e. in terms of the highest-order concept of the self-consciousness property. And of course this is not to say that we cannot make these things rationally intelligible at all.

Compare this innocuous regress of Kantian self-consciousness with the rather more vicious regress of orders of Humean desire implied by Frankfurt's notion of higher-order desires

that evaluate rationally the first-order desires of the self.[29] I critiqued Frankfurt's view in Volume I, Chapter VIII. 2. There I argued that this notion is unsuccessful in providing terminating criteria of rational self-evaluation because we make a commitment to any such n-order desire as authoritative arbitrarily, by fiat. By contrast, the regress of Kantian self-consciousness successfully provides authoritative terminating criteria of rationality intelligibility for all our experiences, including our desires – not by fiat, however, but rather by definition: If something is not recognizable to me as an object of my experience, then however else, and to whomever else it *is* identifiable, it cannot be rationally intelligible *to me.*

Second, that we must be able to make things rationally intelligible as objects of our experience explains why no concept *other* than that of the self-consciousness property could be the highest-order one within an agent's perspective – as we have just seen with the supposedly higher-order concept of an event in the world. Recall that Kant thought that the holistic regress necessarily terminated in *three* highest-order concepts, those of God, freedom, and immortality. He thought these concepts were determined by the most basic categories in terms of which we make anything rationally intelligible to ourselves (1C, A 643/B 671 – A 644/B 672, and especially 1C, A 651/B 679); and that these, in turn, were derived from, respectively, the disjunctive, hypothetical, and categorical judgment forms and syllogisms of theoretical reason (1C, A 321/B 378 – A 338/B 396).

Many of Kant's assumptions here can be called into question. But the important one is the one that Kant himself, in his later writings, also came to question, namely that there are three distinct and irreducible syllogistic forms to begin with.[30] It now seems clear that the categorical syllogism is fundamental: This is the one by which we both directly and inferentially identify things and properties as being of certain higher-order kinds. But we have just seen that the *only* highest-order property in terms of which we can ultimately identify things that is consistent with our preservation of the rational intelligibility of those things *to us*, and so of our own agency, is the property of the thing as an object of our experience, i.e. the property of self-consciousness. This is why I maintain that Kant was right about the holistic regress, but wrong about the highest-order concept to which it inevitably leads us.

But it may be objected, finally, that there are many things that are rationally intelligible to me that I cannot possibly identify as objects of my experience: electrons, for example; or irrational numbers. However, this objection depends on conflating the all-inclusive property of something's being an object of my experience with the narrower ones of its being an object of my sensation, perception, or intuition. As I have tried to develop the concept, the objects of my experience are inherently theory-laden. The theory that loads them may be more or less sophisticated, and the things themselves more or less perceptually concrete. If the higher-order concepts that make things rationally intelligible to me are or include those of theoretical physics, then the entities described by theoretical physics are objects of my experience just in case they, in turn, satisfy the criterion of vertical consistency relative to some of the lower-order properties as,

for example, those that describe a cloud chamber and the perceptually observable processes that occur within it.

In *The View From Nowhere*[31], Thomas Nagel argues that the capacity for transpersonality (my term, not his) inherent in our ability to ascend to higher, more objective and external levels of conceptual abstraction, within which one's subjective, internal perspective may be situated as one among many, exposes us to the purportedly inevitable danger of what he calls *double vision*: a split in the self caused by a purportedly insoluble subjective-objective conflict between the way the world appears to us (however sophisticated and theory-laden that appearance may be), and the viewpoint "from outside," *sub specie aeternitatis*, relative to which we are forced to a principled skepticism concerning the veracity and comprehensiveness of any and all of our beliefs.[32] I conclude this chapter with two brief remarks on Nagel's thesis.

First, the process of ascending to higher orders of conceptual abstraction, as I and Nagel conceive it, contains no *inevitable* subjective-objective conflict. A rationally integrated agent is one who is cognizant both of the fine-grained, perspectival singularity of concrete particulars, and simultaneously of their broader significance as instances of concepts and principles; both of their personal associations and of their impersonal implications. This is a form of cognition in which concrete particulars are viewed as broadly meaningful – or even profound – precisely because they retain both their perspectival particularity for the cognizer and also their function as exemplars of more abstract and impersonal concepts and principles. In contrast with Nagel's double vision, call this form of cognition *transpersonal integrity*.

Double vision, as Nagel describes it, occurs only when transpersonal integrity is violated; that is, when we fail to attend simultaneously to lower-order properties and the higher-order ones we predicate of them. And while we might fail to *attend* to both, the holistic regress implies that we could not fail at least implicitly to *conceive* both, without eradicating both higher- and lower-order properties altogether. We fail to attend to higher-order properties when we perceive some event or state of affairs as a more or less concrete occurrence, but as lacking the significance attention to its more abstract properties might impart. So, for example, one might perceive an acquaintance's offhand remark as factually false, failing to understand it as a joke – and indeed a heavily irony-laden one at that; and so set about earnestly correcting his factual mistake, angrily reprimand onlookers for laughing at him for making it, be mystified by their subsequent condescending attitude toward oneself, and so on. We criticize such an agent as too *literal-minded*, meaning by this that the person fails to grasp the larger or contextual meaning of an event or state of affairs.[33] If he fails to grasp it even after explanation, or fails to grasp the larger meaning of too many such events or states of affairs, we rightly infer that he is missing a chip. The holistic regress implies that such an agent will have failed to perceive such particular events and states of affairs accurately in the first place.

By contrast, we fail to attend to the lower-order properties when we lose a sense of the concrete particularity of the event or state of affairs, seeing it as nothing but an instantiation of the abstract property to which one attends. Some of us, upon realizing the problem of other

minds, really do view others as coats stuffed with straw; some of us, upon learning the theory of natural selection, view others as entwined digestive and central nervous systems encased in flesh; and some of us, upon learning theoretical physics, view all three-dimensional objects as mere perturbations in the force field. And some of us develop such perspectives on the surrounding world without the promptings of philosophical sophistication at all. The phenomenon is called *depersonalization*, and it is a familiar form of psychopathology to which different individuals are susceptible to different degrees, for different reasons, and at different times. To depersonalize an experience is to erase its concrete, personal and self-referential component. Depersonalization undermines the transpersonal integrity of an agent's perspective by attenuating the connection between the higher-order, more inclusive concepts that systematize our perspectives, and the particular set of concrete perceptions, feelings, thoughts and intuitions that ground them – and, in so doing, make our perspectives our own.

Here, too, there are of course degrees. At one end, there is the occasional, momentarily disorienting view of other people as puppets or animals, in which one withdraws the conceptualization of their behavior as personally and subjectively meaningful. At the other, there is the full-grown loss of the sense of individuated self altogether – the loss of the sense that one's perspective is anchored in sentient agency. At this extreme, patients sometimes report, for example, not really feeling the physical pain they recognize themselves to be having; or of lacking the motor sensations of walking although they recognize their locations to be changing; or of "life" in general being literally meaningless. In all of these cases, depersonalization undermines voluntary agency by distorting its objects, disconnecting the causal efficacy of its motives, stifling its sensory impact, or deflating its purposes.[34] All such perspectives have in common the temporary or permanent loss or distortion of the significance of the concrete, and a corresponding dysfunction of the integrative role of theoretical rationality in grasping it. It would be a bad mistake to explain the pathology of depersonalization as inherent in the very cognitive capacity whose normal functioning averts it.

Why our attention is sometimes diverted from the abstract to the concrete, and other times from the concrete to the abstract is a matter for empirical psychology and contingent empirical circumstance to explain. Accordingly, we must then explain the phenomenon of double vision in similarly contingent psychological terms, by finding out why some of us are so quick to abandon or ignore the concrete, particular experiences that necessarily ground our more abstract viewpoints whereas others are not.

Second: If, as I have argued, the self-consciousness property is inevitably our highest-order concept, no matter how high we ascend in orders of conceptual abstraction, then transpersonal integration of the subjective with the objective viewpoints is always a psychological possibility. For even the viewpoint "from outside," *sub specie aeternitatis*, and the principled skepticism it engenders, is itself an object of our experience. An experience can be impartial, impersonal, objective, and abstract – and still be one's own. An agent's perspective that bears these characteristics is the perspective of transpersonal rationality.

Endnotes to Chapter II

[1]David Lewis makes a valiant attempt to replace propositions as objects of intentional attitudes with self-ascriptions of the corresponding properties in "Attitudes *De Dicto* and *De Se*,"in *Philosophical Papers, Volume I* (New York: Oxford University Press, 1983). Also see Brian Loar, "The Semantics of Singular Terms," *Philosophical Studies 30* (1976), 353-77; and John Perry, "The Problem of the Essential Indexical," *Nous 13* (1979), 3-21. The following arguments apply to propositions whether analyzed in terms of states of affairs, possible worlds, or situations, provided only that they are sentential in form.

[2] I discuss Hume's conception of theoretical reason in Volume I, Chapter XIV.

[3]I speak of sentential propositions rather than sentences in order to avoid the implication that one must have or use a language in order to be theoretically rational. The significance of this will become clearer in Chapters II and III.

[4] See Robert Brandom, *Making It Explicit: Reasoning, Representing, and Discursive Commitment* (Cambridge, Mass.: Harvard University Press, 1994). Henceforth references to this work are paginated in the text in parentheses preceded by "MIE." Also see his *Articulating Reasons: An Introduction to Inferentialism* (Cambridge, Mass.: Harvard University Press, 2001).

[5] See Jerry Fodor and Ernest Lepore, *The Compositionality Papers* (New York: Oxford University Press, 2002). In taking the position that only thought and not language is compositional, Fodor's "Language, Thought and Compositionality" (*Mind & Language 16*, 1 (February 2001), 1-15) raises a new host of questions that are beyond the scope of this discussion, so I leave it aside for present purposes.

[6] Michael Dummett, *Frege's Philosophy of Language* (New York: Harper and Row, 1973), 417; quoted in Brandom MIE 339 with added italics.

[7] This gloss on Brandom simplifies his proposal with regard to terminology and scope, but does not affect its import for present purposes.

[8] My account is compatible with Béatrice Longuenesse's more detailed and scholarly treatment in her *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). See especially Chapters 1 and 2.

[9] I refer to *intentional attitudes* rather than *propositional attitudes* in anticipation of the arguments to come. Briefly, these conclude that sentential propositions are not the only intentional objects we have these attitudes toward.

[10]See Joel Feinberg, "Action and Responsibility," in *Doing and Deserving* (Princeton, N. J.: Princeton University Press, 1970), p.134 ff; also John Austin, "A Plea for Excuses," *Philosophical Papers* (Oxford: Clarendon Press, 1961), p.149.

[11] I shall not address here the standard questions about whether the "is" in (7) is really the "is" of predication or the "is" of identity.

[12] See Ernest G. Schachtel, "On Memory and Childhood Amnesia," *Psychiatry 10* (1947), 1-26; and Ulric Neisser, "Cultural and Cognitive Discontinuity," in T. E. Gladwin and W. Sturtevant, Eds., *Anthropology and Human Behavior* (Washington, D. C.: Anthropological Society of Washington, 1962).

[13] *Methods of Logic*, Third Edition (New York, N. Y.: Holt, Rinehart, and Winston, 1972), Chapter 14.

[14]Kant was wrong to drop this useful notion from the B Edition, since it captures the case of recognizing something as an object independently of knowing what kind of object it is.

[15]See D. M. Armstrong, *Belief, Truth and Knowledge* (London: Cambridge University Press, 1973), Chapter 5.

[16] I discuss the interpretation of these passages, and Kant's view of reason more generally elsewhere. See my "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, which itself previews *Kant's Metaethics: First Critique Foundations* (in progress).

[17] I discuss Kant's view of basic concepts as rule-governed, judgmental functions for synthesizing representations of concrete particulars into intelligible categories of experience at *ibid*.

[18] Elsewhere I show how we can understand this relation without imputing to Kant an objectionable or exotic metaphysics of the sort for which Kant is, in many circles, infamous. See *ibid*.

[19] but not entirely; see Section 6 and Chapter III below.

[20] Here I make some very shaky assumptions, which I do not really believe, about statistical "normalcy," when in fact these assumptions must be strictly relativized to the economically privileged classes of political stable, industrially developed countries. Globally, these of course comprise a distinct statistical minority. Hence I assimilate these assumptions to the idealizations otherwise deployed in this first Part of the discussion.

[21] See Robert Paul Wolff, *Kant's Theory of Mental Activity* (Cambridge, Mass.: Harvard University Press, 1968) for a discussion.

[22]See, for example, P. F. Strawson, *The Bounds of Sense* (London: Methuen, 1968), Chapter II.2. In hindsight Kant himself grudgingly admits that hypothetical and disjunctive syllogisms contain the same "matter" as the categorical judgment, but refuses to budge on their essential difference in form and function. See Kant's *Logic*, L, Paragraphs 24-29, 60, Note 2, especially Paragraphs 24, Note – 25; and Paragraph 60, Note 2.

[23] In Section 6, below, I offer some further reasons for preferring my neo-Kantian revision to Kant's original formulation. Its application within a decision-theoretic conception of preference in a variable term calculus is discussed in Chapter III.9, following.

[24]This thesis is elaborated in the contemporary context by Gerald M. Edelman, *Neural Darwinism: The Theory of Neuronal Group Selection* (New York: Basic Books, 1987) and *The Remembered Present: A Biological Theory of Consciousness* (New York: Basic Books, 1989). See the review of Edelman and others by Oliver Sacks in "Neurology and the Soul," *The New York Review of Books XXXVII*, 18 (November 22, 1990), 44-50.

[25]Also see Roderick Chisholm's *Person and Object: A Metaphysical Study* (La Salle, Ill.: Open Court, 1976).

[26]A variant on this criticism is made by E. Moody of Porphyry's interpretation of Aristotelian logic (see E. Moody, *The Logic of William of Ockham* (New York: Russell and Russell, 1965), 70-75). I am grateful to Thomas McTighe for directing me to this source.

[27] The ideas in the following paragraphs benefited greatly from careful study of Joel Feinberg's "The Idea of a Free Man," in *Rights, Justice, and the Bounds of Liberty* (Princeton, N.J.: Princeton University Press, 1980).

[28] For these reasons, I take issue with Bernard Williams' claim that "When I think about the world and try and decide the truth about it, ... I make statements, or ask questions, ... [which] ... have first-personal shadows, ... [b]ut these are derivative, merely reflexive counterparts 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" (*Ethics and the Limits of Philosophy* (Harvard University Press, 1985), p. 67). If I did not occur in such statements in the role of one who had this thought, I would be unable to act on any thought I had. So I think Williams is too quick to differentiate the "I" of theoretical deliberation as necessarily impersonal from the "I" of practical deliberation as necessarily personal. I argued in Volume I, Chapter VIII. 3. 2 that impersonality in deliberation is a function of purely psychological factors, not moral or philosophical ones.

[29] This notion is developed in Harry Frankfurt's "Freedom of the Will and the Concept of a Person" (*The Journal of Philosophy LXVIII*, 1 (January 1971), 5-20).

[30] See Kant's *Logic*, *op. cit.* Note 21, L, Appendix to the Introduction, Ak. 86-87; Pars. 24, 73. Also see the *Prolegomena*, P, Ak. 325, fn.

[31](New York, N.Y.: Oxford University Press, 1986).

[32] *Ibid*., 74-89.

[33] Another, particularly baroque variation on the bullying tactics described in Chapter I.3 is to refuse an interloper entry to a socially and linguistically defined philosophical community by refusing to recognize as a joke an utterance clearly meant to be one; and instead performing on it a Philosophy 101-style linguistic analysis that earnestly refutes its semantic and metaphysical presuppositions. We might describe this as *faux-literal-mindedness*, a stance intended to reprimand the interloper for presuming familiarity with the reigning linguistic conventions.

The effect, of course, is to call into question the viability of those conventions and to reconsider one's interest in joining the community defined by them (to paraphrase the Rolling Stones' famous dictum).

[34]For some fascinating first-person accounts that bear more than a passing resemblance to Nagel's notion of double vision, see John Custance, "The Universe of Bliss and the Universe of Horror: A Description of a Manic-Depressive Psychosis," especially pp. 58-60; Marguerite Sechehaye, "Excerpt from *Autobiography of a Schizophrenic Girl*," especially pp. 166-169; Eugene Meyer and Lino Covi,"The Experience of Depersonalization: A Written Report by a Patient," especially pp. 255 and 258; and William E. Leonard, "Excerpt from *The Locomotive God*," especially p. 312; all collected in Bert Kaplan, Ed. *The Inner World of Mental Illness* (New York: Harper and Row, 1964).

Chapter III. The Concept of a Genuine Preference

Chapter II analyzed one particular intentional attitude – intention – in order to introduce a definition of subsentential constituents that includes among them nonsentential intentional objects. It also proposed some basic notational revisions to classical logic that enabled both their symbolic formulation and their subordination to its basic requirements. Preferences, like intentions, are intentional attitudes. They are also like intentions in being able to take nonsentential intentional objects that cannot be reduced to sentential formulation. My particular interest is in preference as this term is used in formal decision theory, to denote objects of rational choice. These are the objects that enter into pairwise comparisons and linear and nonlinear orderings.

As we have seen in Volume I, Chapters III through VI, something can be an object of choice without being an object of desire. In fact, an object of intention can be, and usually is, such an object. This particular intentional attitude, which anchored the discussion of Chapter II, is an example – and for a Kantian conception of the self, the most important kind of example – of a preference that bears no necessary relation to desire. Therefore, the denotational scope of the term "preference" is not restricted to desire, or happiness, or satisfaction. Some preferences are intentions, some are resolves, some are desires, and some are ground projects or conceptions of the good. The term "preference" as I use it here covers all such nonsentential intentional objects. It would also cover sentential intentional objects of such attitudes, but those do not require our attention in this project.

My aim is now to show in greater depth that these nonsentential preference objects similarly can be brought within the purview of classical logic's consistency requirements. Thus these notational revisions enable us to take up and resolve some of the issues left hanging in Volume I, Chapter IV.2 – 3. There I promised to explain the sense in which the Ramsey-Savage notion of transitive consistency is a special case of a more comprehensive principle of logical consistency; and therefore the sense in which formalizations of the utility-maximizing model of rationality similarly instantiate a more comprehensive Kantian model. In this chapter I attempt to make good on that promise. I show that preference objects, including but not limited to those which maximize the satisfaction of desire, must meet consistency requirements not to be found within the scope of the utility-maximizing model of rationality itself.

In Section 1 I pose the problem for the utility-maximizing model of rationality: essentially that canonical decision theory lacks the technical and formal resources to state in what, exactly, the "inconsistency" of a cyclical ranking consists. I argue that the apparent insolubility of this problem lies in the inadequacy of canonical decision-theoretic notation, which since its historical inception and regardless of its other innovations has persistently concealed its own intensionality. Section 2 makes the case that the way to solve this problem is to rethink and revise the notation. Section 3 derives some criteria a successful decision-theoretic notation would have to meet. Section 4 critically evaluates one possible proposal for revising the notation, and rejects it

on the grounds that it fails these criteria. It does, however, suggest additional criteria that also need to be met. I suggest that these require a variable term calculus that integrates the decision-theoretic concept of preference into the language of classical predicate logic.

Section 5 introduces some basic notational revisions that define the proposed variable term calculus; demonstrates their fidelity to central conventions of both truth-functional and predicate logic on the one hand, and traditional decision-theoretic notation on the other; and gives some examples of how to navigate with these revisions. The most extended of these examples is the reduction of Luce and Raiffa's indifference relation to my notion of Epicurean indifference. Section 6 then applies the calculus to the analysis of two alternative views. First it compares my analysis of indifference to Mark Kaplan's account of rational indecision. Second, it outlines the rudiments of an "occasional" truth-functional analysis for subsentential constituents, and applies that analysis to the Jeffrey-Bolker representation theorem with respect to the thesis – itself a refinement of Ramsey's reasoning – that indifference is an equivalence relation fully adequate to the extensional work such a relation must do. Section 7 introduces four of my five suggested formal criteria a logically consistent series of pairwise comparisons must meet. Here I demonstrate how, using the conventions of predicate logic rather than traditional decision-theoretic notation, we can dispense with talk about "imposing axioms" that – as I showed in Section 1 – gets us into trouble in the first place. Section 8 introduces some further notational revisions necessary in order to formulate the fifth criterion, i.e. ordinality, in terms of the variable term calculus; and introduces and discusses that criterion. Section 9 contrasts the notion of subsentential predication developed in Section 8 with De Jongh and Liu's structurally similar analysis of strict preference in terms of constraint-predicates derived from optimality theory. Section 10 embeds the resulting, logically consistent conception of genuine preference within the comprehensive rationality constraints on coherent experience in general – i.e. horizontal and vertical consistency – which I introduced in the preceding chapter.

Section 11 then applies the resulting Kantian model to the problem described in Section 1, and demonstrates that this model solves it. Since a noncyclical preference ordering is nonvacuous, this alternative, more comprehensive Kantian model of rationality avoids the vacuity of the unreconstructed utility-maximizing model, by demonstrating in what precise and formal sense a cyclical preference ordering is logically self-contradictory.

1. A Problem about Cyclical Inconsistency

We saw in Chapter IV.2 – 7 of Volume I that sequential pairwise comparisons among a given set of preference alternatives F, G, H, … can result in a cyclical ranking

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

and that by invoking the transitivity axiom

(T) if F>G and G>H then F>H,

we can derive from this the "inconsistency" of

(1) F>H and H>F.

I suggested, but did not defend the suggestion that the only viable concept of inconsistency we have is the one we find in classical logic,[1] i.e. violation of the law of noncontradiction. Without this concept (or one comparably sophisticated), it will not do simply to *call* (1) inconsistent, without explaining specifically in what sense it *is* inconsistent. We can agree that there is something wrong with an ordering that includes both F>H and H>F; and it is very tempting to say that they "in some sense contradict" each other. However, we also saw that neither orthodox decision-theoretic notation nor classical logic notation seems to contain the resources for symbolizing in what the "inconsistency" consists. For just as (1) becomes

(2) P.Q

in standard sentential logic, or

(3) $(\exists f)(\exists h)(Pfh.Phf)$

or something equally unilluminating in standard quantificational logic, similarly (T) becomes

(4) (P.Q) → R

in standard sentential form, or

(5) $(g)(\exists f)(\exists h)(Pfg.Pgh \rightarrow Pfh)$

in standard quantificational logic. It is irritating not to be able to symbolize formally the logical inconsistency involved in (1) because this inability undermines the shared intuition that there is one.

The natural response to this irritation – indeed, the standard move – is to insist on the distinction between normative decision-theoretic axiom systems and the descriptive empirical choice behavior that to varying degrees may or may not approximate them. By imposing decision-theoretic versions of certain axiomatic conditions derived from classical logic such as transitivity and asymmetry on the behavior of an ideally rational chooser, we exclude cyclical rankings from the scope of the normative system. However, denying the existence of a cyclical ranking within a normative decision-theoretic axiom system does not eliminate it in reality. Relative to that wider empirical reality, there is no detectable logical inconsistency between (C) and (T), hence none between the terms of (1).

We see here a significant disanalogy between classical logic on the one hand, and formalized decision theory on the other. In a classical axiom system, imposing axiomatic conditions such as transitivity and asymmetry rules out logical contradiction in a way that explains and effectively predicts with 100% accuracy the corresponding absence of logical contradiction to be found anywhere in empirical reality. That is, the limits of logical possibility defined by the system mirror the limits of logical possibility found in reality. In a normative decision-theoretic axiom system, by contrast, imposing the decision-theoretic analogues of transitivity and asymmetry exclude cyclical rankings from the system without excluding them from the wider empirical reality in which that system is situated.

Thus the standard move, of imposing decision-theoretic axioms in order to rule out cyclical “inconsistency,” does not dissolve the irritation because normatively excluding the “inconsistency” a cyclical ranking represents does not eliminate it. The fact of the matter is that

some subjects do exhibit cyclical selection behavior. Unlike a logical inconsistency, a cyclical ranking seems to remain a *logical possibility* relative to decision theory, despite the imposition of classical logic-like axioms that normatively exclude it.

This fact by itself illuminates the background context of classical logic relative to which decision-theoretic axioms must be interpreted. What enables us to recognize a cyclical ranking as a logical possibility outside a normative decision-theoretic formalization is the more inclusive, background constraints of classical logic that define what a logical possibility is. These more inclusive background constraints define the outer limits not only of axiom systems of classical logic. As I tried to show in the preceding chapter, they define the outer limits of our experience of empirical reality as well. We do not need to study an axiom system in order to be quite certain that it is not logically possible for both P and not-P to be true at the same time in the same respect, because the horizontal and vertical consistency of our experience itself ensures this. Since (1) does not assert that both F and not-F, (1) appears to be logically possible. What we lack is a decision-theoretic version of this "reality test" to establish symbolically what we intuitively already know with equal certainty: that in fact it is no more logically possible for an agent to prefer F to H and H to F at the same time in the same respect than it is for both P and not-P to be true at the same time in the same respect. The horizontal and vertical consistency of our experience excludes both. A decision-theoretic notation that enables us to symbolize the former as an instance of classical logic's symbolization of the latter does not seem too much to ask.

The disanalogy between classical logic and normative decision theory carries through to agent behavior, where it continues to work to the disadvantage of the latter. When a subject sequentially asserts that P and then asserts that not-P, we credit her with intertemporal logical consistency by inferring, in accordance with the principle of charity, that she has changed her mind. Or we may infer – less charitably – that she is speaking irrationally. Neither the authority, the legitimacy, nor the scope of classical logic are threatened by these inferences, because the constraints on reality that classical logic mirrors themselves force the conclusion that the agent must have misspoken – rather than that the logic must be revised.

By contrast, we have already seen in Volume I, Chapter IV.2 – 3 that when an agent sequentially selects F>G, G>H, and H>F, we have two analogous options, neither of which is comparably benign in its effects. First, we can save the rationality of the ranking by similarly applying the principle of charity – which immediately uncovers the vacuity of the underlying principle of utility-maximization (U). Second, we can attempt to make the inference to irrationality – only to be thwarted once again by the "universality" (actually the promiscuity) of (U)'s scope of application: if this is the ranking that maximizes utility for this agent, then there is nothing irrational about it. Both of these options do threaten the authority, legitimacy and scope of (U) in its unreconstructed form, because (U) fails to mirror any particular reality constraints. This is what it means to call (U) vacuous, or promiscuous in its scope of application; and what makes the case for revising its logic.

Under these circumstances it may well seem that the only remaining alternative is the standard move: to acknowledge the empirical reality of the cyclical ranking, then rule it out normatively by imposing the axioms – which secures its *de jure* irrationality while preserving it *de re* as a seeming logical possibility nevertheless. But this is not the only remaining alternative, and it is not a good one. I suggested in Volume I, Chapter IV.2 – 3 that the main effect of imposing axiom conditions so as to restrict the scope of application of (U) to the narrowly normative is that "consistency" is preserved at the expense of universality of application, whereas removing the conditions preserves universality – to the point of vacuity – at the expense of "consistency".

The problem is exactly analogous to that which I then examined in Volume I, Chapters IX – XI, encountered by Humean metaethical views that impose normative conditions on instrumentally rational choosers: the more such conditions are imposed, the more the outcome is restricted to the normatively moral at the expense of its objective (or intersubjective) validity; whereas lifting the conditions increases the objective validity of the choosers' choice at the expense of its normative morality. The reason the failure takes the same form in both cases is that the relation between the two terms in each choice is exclusive rather than implicative: Just as normative morality and objective validity are mutually exclusive in an instrumentalist justificatory scheme, so, too, are normative consistency and universality in a utility-maximizing model of rationality more generally. Indeed, the former is an instance of the latter.

Excluding cyclical rankings through the imposition of normative conditions alone drives an unnecessary wedge between our conceptions of what rationality requires, what logic demands, and what reality permits. The scope of the logically possible has an authority backed by what reality permits. That authority is not superseded by the more restricted scope of the rationally admissible that the utility theorist stipulates in order to exclude the reality of cyclical rankings from its normative purview. On the contrary: that reality undermines the authority of the normative purview its exclusion helps to define.

Because the normative-descriptive distinction does not solve the problem of how to parse the cyclical “inconsistency” of (1), it is tempting to draw a cruder distinction, between classical logic and decision theory simpliciter; i.e. to contend that decision theory applies to a radically different “realm” – perhaps the “realm” of the free will – from that of classical logic, in the same ways – and perhaps for the same reasons – that practical reason is inherently different from theoretical reason and action is inherently different from thought. I reject the distinction between practical and theoretical reason explicitly in Chapter V below. I call these *pseudo-Kantian* distinctions because I do not believe this was Kant’s view, although it is often attributed to him. None of these pseudo-Kantian distinctions is ultimately convincing. Action just is intentionally conceptualized behavior of a goal-oriented kind, and so presupposes thought. Practical reasoning just is an application of theoretically rational rules of causal inference to the special-case event of intentionally conceptualized behavior of a goal-oriented kind. Then if decision theory is a formalization of practical reasoning, then it is a special case of the classical logic that formalizes theoretical reasoning. The question is how to show this.

Now it might be argued that decision-theoretic formalizations are best compared not with classical logic, but rather with intensional logics of belief, in which it is logically possible for a subject to have logically contradictory beliefs *bP* and *b~P* simultaneously. But first, we saw in discussing Ramsey's value axioms in Volume I, Chapter IV.2.1 – 2 that neither orthodox decision-theoretic notation nor decision-theoretic idiolect recognizes an intensional component to preference rankings. The use of the passive voice, as in "F is preferred to G" conceals any that might be there, and gives the interpretation of decision-theoretic symbols a strong extensional cast. Second, even if intensional logics of belief were the correct comparison, the similarity would break down at the next step. For intensional logics of belief have to acknowledge the logical possibility of contradictory beliefs as admissible within the system in order to preserve the logically consistency of the system itself; whereas both classical and decision-theoretic axiom systems make strong claims to exemplify in what consistent theoretical and practical reasoning respectively consist.

Nevertheless, the issue of intensionality first encountered in discussing Ramsey's value axioms is unavoidable. For – as I now argue – the reason we cannot formalize cyclical inconsistency within the constraints of orthodox decision theory is because orthodox decision theory treats as extensional connectives what are in fact intensional operators buried in declarative sentential propositions.[2] Here is an argument that purports to derive a straightforward logical inconsistency from the conjunction of (C) and (T). Reading the weak preference relation F ≥ G as "F is preferred or indifferent to G," define a strong preference relation F>G in terms of it such that

(6) F≥G =df. weak preference

(7) F≥G . ~G≥F =df. strong preference (F>G)

(8) (F≥G . G≥H) → F≥H =df. transitivity for weak preference

(9) [(F≥G . ~G≥F) . (G≥H . ~H≥G)] → (F≥H . ~H≥F)
=df. transitivity for strong preference

(10) (F≥G . ~G≥F) . (G≥H . ~H≥G) . (H≥F . ~F≥H)	cyclical ordering
(11) (F≥G . ~G≥F) . (G≥H . ~H≥G)	(10)
(12) (F≥H . ~H≥F)	(9), (11)
(13) (H≥F . ~F≥H)	(10)
(14) F≥H . ~F≥H . H≥F . ~H≥F	(12), (13)[3]

But this derivation is not as straightforward as all that. Its truth-functional connectives connect neither standard sentential propositions nor sentences that can be replaced by extensional sentence letters P, Q, R, …. Here is what happens when we try:

(6′) P

(7′) P . ~Q

(8′) (P . R) → S

(9′) [(P . ~Q) . (R . ~T)] → (S . ~U)

(10′) (P . ~Q) . (R . ~T) . (U . ~S)

(11′) (P . ~Q) . (R . ~T)	(10′)
(12′) (S . ~U)	(9′), (11′)
(13′) (U . ~S)	(10′)
(14′) S . ~S . U . ~U	(12′), (13′)

(9′) – (14′) constitute a valid derivation of a logical contradiction, all right; but not from the conjunction of (C) and (T). (C) and (T) have disappeared, buried in the extensional formulations of (9′) and (10′). This standard truth-functional derivation fails to demonstrate the logical inconsistency of a cyclical ranking because by substituting extensional sentence letters for the variable terms of (6) – (14), it deletes the extra, quasi-logical connective "≥", and thereby conceals the signs of transitivity, cyclicity, and the problems that arise from conjoining them.

Clearly, we are not in Kansas anymore. The classical Boolean connectives ".", "~" and "→" relate extensional sentential propositions. By contrast, "≥" and ">" as imported into decision theory from mathematics – and, for that matter, the "r" and "p" that often replace them in more recent treatments – in fact are not really connectives at all. "≥" and ">" (and "r" and "p") instead express intentional attitudes toward pairs of intentional objects. (6′) – (14′) shows that we are not free simply to add "≥" (or "r") on to the list of classical Boolean connectives and perform the same sorts of logical operations on it as we are used to doing on them. Hence we similarly cannot perform the standard logical functions on the variable terms related by "≥" plus the classical Boolean connectives in (6) – (14) either, because all such terms embed an extra, intentional operator within their logical substructure, and all make assertions the intensional content of which consequently resists the degree of intersubstitutability that the Boolean connectives require.

But when the intensional structure of these assertions is exposed, further problems ensue. Take (9), transitivity for strong preference. For what kind of chooser does (9) hold true? Not for an actual chooser, since (9) is not a prediction. And not for an ideally rational chooser under conditions of uncertainty, since in that case the chooser's preferences are not epistemically transparent (for example, from the fact that I prefer peaches to pears and pears to cherries, does it follow that I prefer peaches to cherries? It is hard to see why it should). Might it hold true for an ideally rational chooser S under conditions of full information? It seems not. (9) can be paraphrased in a way that exposes its intensional structure as follows:

(9″) If it is the case that

(9″.1) S prefers F to G or is indifferent between them, and

(9″.2) it is not the case that S prefers G to F or is indifferent between them;

and that

(9″.3) S prefers G to H or is indifferent between them, and

(9″.4) it is not the case that S prefers H to G or is indifferent between them;

then it is the case that

(9″.5) S prefers F to H or is indifferent between them, and

(9″.6) it is not the case that S prefers H to F or is indifferent between them.

Note that indifference does not satisfy symmetry in either of the antecedent conjuncts of (9″), or in its consequent.

(9″), in turn, would seem to imply that if it is the case that

(9″.7) S is indifferent between F and G, and

(9″.8) S is indifferent between G and H;

then it may be the case that

(9″.9) S prefers F to H, and

(9″.10) S is indifferent between F and H.

This implication of (9″) seems intuitively self-contradictory. Hence (9) cannot be presumed to hold, even of an ideally rational chooser under conditions of full information.

The prima facie plausibility of (9) – and the derivation (6) – (14) – depends on concealing their intensionality by repackaging what is in fact an intentional operator – the preference operator – as a quasi-logical connective of mathematical ancestry. But we have just seen that in decision theory, neither "$>$" nor "$\geq$" are genuine relational connectives, whether logical or quasi-logical, because they do not connect extensional terms. They are rather symbolic expressions of intentional operators that denote certain of a subject's intentional attitudes – namely, preference and weak preference respectively – toward pairs of intentional objects – namely preference alternatives. So if a rule of transitivity of preference is going to hold in decision theoretic formalizations, the intensional conditions under which it does hold need to be spelled out.

I believe these conditions can be spelled out consistently with showing the sense in which (1) is logically self-contradictory, and so the sense in which classical logic provides the "reality test" that authorizes the decision-theoretic rejection of cyclical preference as logically contradictory. I address this task in Section 11, below. Of course that (1) can be shown to be logically self-contradictory does not imply the logical *impossibility* of cyclical *selection behavior*, any more than it would the logical impossibility of self-contradictory speech behavior. What it does imply is that choice behavior is just as much subject to the consistency requirements of classical logic as speech behavior, in both normative and descriptive systems; and so that the utility-maximizing model of rationality is similarly subject to a more inclusive, Kantian model of rationality that places classical logic at its base.

2. Savage's Concept of a Simple Ordering Reconsidered

I argued in Volume I, Chapter IV.2.3 that Savage's concept of a simple ordering was insufficient to ensure transitivity of preference through three sequential pairwise comparisons, because his "logic-like" rule of transitivity (T), Section 1 above, is neither among nor implied by the laws of logic. I contended that if something like (T) were a logical axiom, it would assert something close to a conceptual truth about what it means to prefer F to G and G to H. Under these circumstances, to violate (T), as does the cyclical ordering (C), Section 1 above, would be to have no genuine preferences among F, G and H at all. Savage seconds that observation:

> [W]hen it is explicitly brought to my attention that I have shown a preference for F as compared with G, for G as compared with H, and for H as compared with F, I feel uncomfortable in much the same way that I do when it is brought to my attention that some of my beliefs are logically contradictory. Whenever I examine such a triple of preferences on my own part, I find that it is not at all difficult to reverse one of them. In fact, I find on contemplating the three alleged preferences side by side that at least one among them is not a preference at all, at any rate not any more.[4]

But we have seen that since F>H does not *logically imply* ~H>F, it seems both may be true together. Therefore (T) and (C) both may be as well.

However, I also argued that there had to be more to transitivity than this, on pain of a moment-to-moment time-dependency in selection behavior so radical as to undermine the necessary conditions of intentional agency. I argued that a conscious and intentional chooser had to satisfy two necessary conditions:

> (a) she must be able to form and apply consistently through time the concept of a thing's ranking superiority – and therefore some other thing's ranking inferiority – over a series of pairwise comparisons; and
>
> (b) she must remember the relation of the two alternatives she is presently ranking to the third she is not.

I argued that a chooser who satisfied these two conditions was minimally psychologically consistent and therefore would produce a transitive series of preference rankings. Further, I suggested that satisfaction of (a) and (b) expressed the correct but extralinguistic assumption that in selecting F over G at t_1 and G over H at t_2, a chooser is applying a time-independent, logically consistent *rule*, namely the concept of a genuine preference; that, like any genuine concept, the concept of a genuine preference provided a criterion of logical consistency, i.e. that a concrete particular alternative not exemplify both it and its negation at one and the same time and in one and the same respect; and that (T)'s arrow therefore should be understood as expressing the conceptual implication that by ranking F over G and G over H, one *thereby* ranks F over H. I concluded that a chooser who has a genuine preference for F over G at t_1 and G over H at t_2 would be constrained by the concept of a genuine preference to select F over H at t_3. So on this account, (T) implicitly expresses a conceptual truth and (C) is logically inconsistent.

In light of the criteria of horizontal and vertical consistency developed in Chapter II.4, we can now see that (a) and (b) presuppose satisfaction of these criteria. That a concrete particular alternative not exemplify both a concept and its negation at one and the same time and in one and the same respect is the requirement of non-contradiction, i.e. $(x)\sim(Fx.\sim Fx)$, for classical predicate logic. This is the sentential analogue of Chapter II.4.1's proposed requirement of horizontal consistency for variables

(HC) $(\sim\exists x)(x.\sim x)$,

i.e. $(x)\sim(x.\sim x)$. The rule that requires consistent application of concepts to particulars in order to exclude self-contradiction applies, *a fortiori*, to the application of the concept of a genuine preference to the particular choice alternatives offered.

The consistent application of this concept over time also presupposes satisfaction of the criterion of vertical consistency developed in Chapter II.4.2. Applying consistently through time the concept of a thing's ranking superiority, and so of some other thing's ranking inferiority, to a series of pairwise comparisons ((a), above) satisfies (B) of Chapter II.4.2, i.e. that any particular c_i in S is either

(i) an instantiation of some other c_j in S; or

(ii) instantiated by some other c_k in S.

That is, suppose S' is that subset of S comprising the choice alternatives available to her at that moment. Then (a) implies that given any choice alternative F in S' that enters into a pairwise comparison with some other choice alternative G in S', F in that comparison instantiates at least one of two other c_js in S', namely the concept of a thing's ranking superiority or of its inferiority; and similarly for any F, G in S'. So S' is minimally coherent.

In addition, remembering the relation of the two alternatives the agent is ranking to the third she is not ((b), above) satisfies (C) of Chapter II.4.2; i.e. that for any cognitively available particular thing t, there is a c_j in S that t instantiates. That is, if F, G, H, …are choice alternatives in S', then F, G and H are cognitively available things t^1, t^2, and t^3 such that there is at least one c_j in S' that each of t^1, t^2, and t^3 instantiate, namely the concept either of a thing's ranking superiority or of its ranking inferiority; i.e. S' is complete. Then for any pairwise comparison among them, an agent who can form and apply these concepts consistently through time can at any moment in time use both of them to locate any one of these choice alternatives relative to the other. She remembers the relation of the two alternatives she is presently ranking to the third she is not.

Now if (a) and (b) presuppose satisfaction of Chapter II.4.2's (B) and (C), then we should be able to symbolize (T) as an instantiation of Chapter II.4.2's formalized criterion of vertical consistency, i.e.

(VC) $Fa \rightarrow [(x)(Fx \rightarrow Gx) \rightarrow Ga]$.

In fact we can do this, but not until Section 10, below. In order to see how (T) satisfies (VC), we need to forge the tools for seeing (T)'s logical structure more clearly than canonical decision-theoretic notation permits us to do.

In the remainder of this chapter I try to make (T)'s logical status – and so that of the concept of a genuine preference – explicit, by reformulating Savage's concept of a simple ordering in such a way as to bring out (T)'s logical – i.e. horizontal and vertical – consistency and (C)'s violation thereof. To do this we need to re-examine and rethink Savage's now-canonical notation for pairwise comparisons.

Savage's original notation began with the mathematical symbol "≤" defined as the "is not preferred to" relation; stipulated that "of any two acts **f** and **g**, **f** is not preferred to **g** or **g** is not preferred to **f**, possibly both;" defined a simple ordering among a set of elements x, y, z …, related

by "≤" as equivalent to that which in addition satisfied transitivity; and proceeded to derive both the indifference relation

(I) **f** ≤ **g** and **g** ≤ **f**

and the "is preferred to" relation ">" in terms of it.[5] An advantage of Savage's notation, in addition to its elegance, is that the move from expressing (T) in terms of his preference relation

(T') If **f** > **g** and **g** > **h**, then **f** > **h**

to the simple ordering

(O) **f** > **g** > **h**

is quick and obvious.

The intuitive plausibility of Savage's notation depends, however, on regarding preference alternatives as numerically commensurable quantities, i.e. on the plausibility of **f**'s being preferred to **g** because **f** is in some sense *more than* **g**; and **g**'s being preferred to **h** because **g** is in the same sense (whatever that is) more than **h**. But this is too narrow. **f**'s being in some sense more than **g** is only one possible basis on which **f** is preferred to **g**, and not the only or even a necessary basis. For example, a chooser might prefer **f** to **g** because **f** is according to some important criterion *different from* **g**. The "different from" relation would be equally capable of ordering multiple alternatives linearly. For example, **f** might be different from **g** according to one criterion; **g** different from **h** according to a second criterion; and **f** different from **h** according to both criteria. The highest-ranked alternative would be that which is different according to as many important criteria as possible. There is no obvious way to translate this relation into a simple calculus of more and less, since being different according to more criteria does not entail being more different.

But even if **g**'s being preferred to **h** is based on **g**'s being in some sense more than **h**, that may not be the same sense in which **f** is more than **g**. For example, (T') might hold even though a chooser finds **f** more reliable than **g**; **g** more appealing than **h**, and **f** more familiar than **h**. If (T') held, then so would (O). Yet expressing the ordered relation among them using only ">" (or complementarily, "≤") would be not only insensitive but misleading. For it would suggest that **f**, **g**, and **h** were ordered linearly according to some uniform measurable criterion, when in fact there was nothing uniform or measurable about it.

Does (O)'s derivative validity imply that there is, or must be, some overarching, measurable uniformity among **f**, **g** and **h** that ">" captures? No, unless their connection through ">" itself suffices. Here it might be argued that it does not matter on what basis a chooser orders **f**, **g** and **h** so long as she orders and connects them with ">". But either this reduces ">" to an arbitrary mark replaceable by any other one carrying an asymmetric connotation (would "@" do equally as well? – of course not); or else it begs the question as to the adequacy of ">" to capture the concept of a genuine preference. The viability of “>” in orthodox decision theory relies for its credibility on its viability in mathematics, where it functions as a genuine connective that relates extensional numerical entities. But we have just seen in the preceding section that this entirely

credible role for ">" in mathematics does not survive unscathed its appropriation into orthodox decision theory.

Suppose, for example, that Gertrude prefers chocolate ice cream (**f**) to vanilla (**g**) because the chocolate tastes sweeter; vanilla to coffee (**h**) because she prefers the taste of vanilla, although neither tastes sweeter than the other; and coffee to chocolate because she prefers the smoother texture of coffee. In Savage's notation, Gertrude's is a typical cyclical ranking:

(1) **f**>**g** and **g**>**h** and **h**>**f**.

Sen's notation treats Gertrude's choice dilemma similarly, with a bit more substructure:

(2) (**f**≥**g**. ~**g**≥**f**) . (**g**≥**h**. ~**h**≥**g**) . (**h**≥**f**. ~**f**≥**h**).

But Gertrude's is not a typical cyclical ranking. For it is not the case that she prefers chocolate to vanilla, vanilla to coffee and coffee to chocolate *simpliciter*. Rather, she prefers chocolate to vanilla and coffee on grounds of sweetness, and coffee to vanilla and chocolate on grounds of texture. Nor can the apparent inconsistency in Gertrude's ranking be described in conventional propositional terms, even had the derivation in Section 1.(6) – (14) gone through for ordinary cyclical inconsistency. For from substituting (2), above, for 1.(10) in Section 1, the derivation of

(3) Gertrude prefers chocolate and it is not the case that Gertrude prefers chocolate and Gertrude prefers coffee and it is not the case that Gertrude prefers coffee and Gertrude prefers vanilla and it is not the case that Gertrude prefers vanilla

does not accurately describe Gertrude's intentional state. There is no suggestion in the description of the case that Gertrude stops preferring chocolate for its sweetness, vanilla for its taste, and coffee for its texture. She continues to prefer each flavor of ice cream for one of its properties, and also something that is not that flavor for a different property. The apparent cyclicity of Gertrude's preference ranking arises out of her failure to rank independently the relevant properties themselves – sweetness, taste, and texture – of the alternatives she confronts. Neither Savage's nor Sen's notation enables us to do that. I suggest a way to do it in Section 8, below.

We may not be able to capture the myriad subtleties of each and every different chooser's preference rankings. But we do not want to beg any questions about what those subtleties are, as ">" does. So if we want to bring out the logical structure of (T), the streamlined elegance of ">", "≤", and "≥" may need to be sacrificed.

3. Notational Desiderata for Preference Alternatives

Savage chose not to symbolize (T) and (C) within the standard constraints of quantificational logic. However, its notation is adequate for the expression of other relational predicates of an intensional nature. If

(1) Everyone loves something

can be expressed as

(2) $(w)(\exists x)Fwx$,

then

(3) Everyone prefers some alternative

can be expressed as

(4) $(w)(\exists x)Pwx$

But if (4) is a legitimate symbolic expression of (3), then

(5) Everyone prefers some one alternative to some other

can become

(6) $(w)(\exists x)(\exists y)Pwxy$.

Expressing preference relations among alternatives is thus far of a piece with expressing other intensional relations among objects in quantificational notation.

This much alone easily expresses relations among the sentential propositions in which reference to these objects occur as subsentential constituents. But it suppresses the structure of relations among those subsentential constituents themselves. This structure is what distinguishes the preference relation from other triadic relational properties, including both extensional ones, such as being the offspring of one's parents, and other intensional ones such as admiring one's partner's mom. Moreover, we have already seen in the preceding chapter that the requirements of theoretical rationality apply to the subsentential constituent objects of intentional attitudes as well as to the sentential propositions in which those objects are embedded; and that these objects cannot always themselves be expanded into further sentential propositions.

The same considerations apply when the intentional attitude in question is a preference. So, for example, the sentential proposition that

(7) Gladys prefers rice and veggies to stir-fry

is not logically equivalent to

(8) Gladys prefers rice to stir-fry and Gladys prefers veggies to stir-fry,

since Gladys may prefer them to stir-fry only when they are combined. Similarly, that

(9) Alonzo prefers charcoal to lead pencil

is not logically equivalent to

(10) Alonzo prefers charcoal and Alonzo does not prefer lead pencil,

since Alonzo's preference in (9) may depend on being offered a pairwise comparison between them.[6] Such counterexamples are familiar motivators for intensional logics, for example, of belief. Similar counterexamples for preference claims could be given for each of the standard Boolean connectives under conventional, natural-language interpretations. What these counterexamples show is that in order to understand and symbolize appropriately the logic of preference, more of the subsentential structure of preference claims need to be exposed.

Only then can we answer the questions posed above, i.e.

(i) Can [the logical analogues of] F>H and H>F both be true together?

(ii) Can [the logical analogues of] (T) and (C) be true together?

As they stand, (T) and (C) express intentional preference relations between individual alternatives F, G and H. (T) and (C) also express seemingly straightforward truth-functional

relations among such sentential propositions as "F is preferred to G" and "G is preferred to H," in which these alternatives are intentionally related. In order to answer (i), we need a notation that can express the difference in intentional status between F and H at least as well as Savage's does. In order to answer (ii), we need a notation that also can reflect the sentential relationships among conjuncts, antecedents, and consequents in (T) and (C) at least as well as Savage's does. An adequate notation will use familiar Boolean connectives under a standard, natural-language interpretation to express both types of relation simultaneously.

4. Some Further Limitations of Standard Quantificational Notation

One possibility that I do not endorse would be to interpret a relational predicate *Fxy* as "*x* selects *F* over *y*," and (T) as

(T^1) $(x)\{[(((Fxg \vee Gxf).\sim(Fxg.Gxf)).\sim Gxf).$
$(((Gxh \vee Hxg).\sim(Gxh.Hxg))).\sim Hxg)] \rightarrow$
$(((Fxh \vee Hxf).\sim(Fxh.Hxf)).\sim Hxf)\}.$

With the addition of an axiom of asymmetry such that

$(x)\ (Fxg \Leftrightarrow \sim Gxf)$, etc.

(T^1) would be equivalent to

(T^2) $(x)(Fxg.Gxh. \rightarrow Fxh).$[7]

This proposes to express the pairwise comparison between F and G as a sentential function

(1) $((Fxg \vee Gxf).\sim(Fxg.Gxf)).\sim Gxf$,

an extended quantificational description of what is involved in selecting between two proffered alternatives. (1) is truth-functionally equivalent to

(2) $Fxg.\sim Gxf$,

which, with the application of the asymmetry axiom, becomes

(3) Fxg.

(T^1)'s second conjunct and consequent similarly can be paired down to essentials.

This notation is pleasing in certain respects. By replacing Savage's preference relation ">" with a chooser *x* in the same location, it preserves through several sentential transformations the same symmetrical placement of alternatives found in Savage's notation that made it an intuitively plausible representation of a pairwise comparison. At the same time, it reformulates this relation in identifiably sentential and quantificational terms. And it offers an expanded quantificational interpretation of Savage's "F>G" that unpacks it sententially.

However, this proposed notation is very counterintuitive in other respects. It is certainly possible to express the preferred alternative in a particular pairwise comparison as a relational predicate, so that the predicate letter changes accordingly with each such ranking. It is also possible to assign the chooser and nonpreferred alternative to variables related by that predicate, so that the preferred alternative is in effect a property that the chooser and nonpreferred alternative are expressed as having. And it is possible to assign to the preference relation itself no

symbolization at all in each of three preference rankings, so that the distinguishing structure of that relation is effectively obscured.

But consider what is thereby lost. There is no syntactic expression of the property, present continuously from one preference ranking to the next, that identifies each as a pairwise comparison, namely the preference relation itself. There is nothing explicitly displayed as a relational predicate that in each pairwise comparison joins chooser with alternatives ranked by it. There are no syntactic resources for distinguishing the preference relation from other triadic relations, whether extensional or intensional – even though, as we have seen in (3. 7)-(3. 10) above, the preference relation has the epistemic opacity that makes it similar to other intentional attitudes and unlike extensional triadic relations. There are no syntactic resources for making the truth functional distinctions explored in (3. 7)-(3. 10) consistent with preserving satisfaction of the transitivity requirement. There is no way of symbolizing preferences that take intentional objects of a non-pairwise comparative type.[8] And there is no way to express syntactically the truth-functional relationship between the preferred and nonpreferred alternatives of a pairwise comparison as itself the disjunctive intentional object of a preference. This seems unfortunate, since to be offered a choice between *x* and *y* would seem at the very least to be invited to choose **either** *x* **or** *y*. One would expect an adequate notation for preference to contain some resources for expressing not only this disjunction – perhaps the suggested notation could be expanded sententially to do that; but also its nonequivalence to the case of either being invited to choose *x* or being invited to choose *y*, in which either invitation expresses a sinister note of coercion.

Because this suggested notation always expresses one alternative of a pair – a different alternative for each ranking – as a relational property, its smallest notational unit is a sentential proposition. In this respect it has some of the same defects as (6) in Section 3. Now I argue elsewhere that Kant believed *all* semantic interpretations to be finally reducible to syntax. Kant's belief may well go too far. But we at least should be able to do better than this. It is not unrealistic to expect an adequate notation for preference to expose enough syntactic substructure to distinguish between one complex intensional object of preference and two or more extensionally distinct preferences. A preference of F to G – itself a subsentential constituent of some sentence – has a syntactical substructure susceptible to analysis in terms that standardize the distinctions – and others like them – to which (3. 7)-(3. 10) call attention.

This substructure can be unpacked in terms borrowed from truth-functional analysis. Interpretation of the subsentential constituent

(4) not stir-fry

cannot be identical to the truth-functional interpretation of

(5) Gladys does not prefer stir-fry,

since (4) is not itself a sentential proposition that can be true or false. The interpretation of (4) will depend instead on the context provided by the sentential proposition in which it occurs – and on the intentional operator that modifies it. In a sentence asserting a preference ranking, (4) will count as a rejection of a proffered alternative. In a sentence asserting a resolve, (4) may count

as the denial of a temptation. In a sentence asserting an intention, (4) may be interpreted as the disclaimer of a goal. In each case, use of the usual Boolean connectives under a natural language interpretation would assign to the "~" the same familiar interpretation of "not"; but in each case the "not" would be nested slightly differently in its context. Similarly with "and", "or", etc.

So we need a notation that can do this. In order to capture the intensional nature of the preference relation, it needs to be able to express both sentential and subsentential relations in familiar quantificational and truth-functional terms. And in order to answer (i) and (ii) of Section 3, we similarly need a notation for expressing recognizably logical relations, not only among sentential propositions, but in addition among the objects assigned to individual variables that are embedded in them as subsentential constituents. I shall call a notation that meets these desiderata a *variable term calculus*.

5. A Variable Term Calculus: Subsentential Applications

In the preceding chapter, I suggested the underpinnings of a variable term calculus in the notation used to limn the concepts of horizontal and vertical consistency developed there. Recall the holistic regress argument from Section 3 of that chapter, and its conclusion that not just sentential propositions, but any rationally intelligible thing or object *t* assigned to an individual variable *a* must satisfy the requirement

(1) $\sim(a.\sim a)$

(in accordance with the conclusions of Section 4 above, read (1) as a sentence fragment that says: "… not both *a* and not-*a* …."). That is, we must conceive *t* as self-identical and so as nonself-contradictory. Recall also that Quine's schematized axioms of identity

(I) $Fy.\ y{=}z.\ \rightarrow Fz$

(II) $y{=}y$

offered a model for schematized axioms of nonself-contradiction, thus:

(I') $Fy.\ \sim(y.\sim z).\ \rightarrow Fz$

(II') $\sim(y.\sim y).$

We also saw that one result of substitution of (I') was

(2) $x{=}y.\ \sim(y.\sim z).\ \rightarrow x{=}z$

from which would follow

(3) $\sim(x.\sim y).\sim(y.\sim z).\ \rightarrow \sim(x.\sim z)$,

which I described as the law of transitivity of nonself-contradiction. I suggested that the requirement of nonself-contradiction among terms and variables might function in proofs, as does the identity sign, either as an inert predicate letter or truth functionally with the insertion of an axiom of nonself-contradiction into the antecedent of the conditional. In turn, (3) implies

(4) $(x \rightarrow y).(y \rightarrow z).\rightarrow(x \rightarrow z)$

and therefore

(5) $(\sim x \text{ v } y).(\sim y \text{ v } z).\rightarrow(\sim x \text{ v } z).$

The value for present purposes of (II') and (3) – (5) is that first, they establish the criterion of horizontal consistency among independent variables, and hence among subsentential constituents. Second, therefore, they demonstrate a way in which the Boolean connectives might function among variables, not only among the quantified sentential propositions that contain them. Basically, these connectives function in a monadic predication of *x* in (II'), and in a dyadic predication of *x* in (3) – (5). These are some of the tautologies that can be imported from the sentential calculus to the variable term calculus I am suggesting.

But the goods to be imported need not be restricted to tautologies. Given certain restrictions to be explicated shortly, the entire truth-functional apparatus of familiar logical connectives, rules of inference, and tests for consistency as well as validity is potentially available. This is fortunate, since whatever (T) is, it is not a tautology. So I shall use the available resources of quantificational notation, in conjunction with the elements of the variable calculus just sketched, in order to fashion a logical analogue of (T).

Let a triadic preference relation *P* for pairwise comparisons be defined as follows. Given variables *w*, *x*, *y* and *z*, let *w* be a chooser and *x*, *y* and *z* be any alternatives – actions, states, events, gambles, compound lotteries, plans, prospects or discrete objects – between which that chooser decides, such that

(P) $Pw(x.\sim y).$[9]

(P) is a sentential function – call it a *strict preference*- [or *P*-] *function* – that states that a conscious and intentional chooser *w* strictly prefers alternative *x* to alternative *y*, i.e. that she selects *x* and rejects *y*. (P) expresses the implicit rejection of the not-preferred alternative involved in all strict preference rankings (if this seems too strong, try convincing the lover you have just jilted that your preference for someone else does not imply your rejection of him, and see how he takes it). So one advantage of (P) is that on an intuitive linguistic level, it captures better what goes on when a chooser makes a pairwise comparison – "this one, not that one" – than the intuitive linguistic reading of Savage's "more than" relation. A disadvantage is that it replaces Savage's asymmetric *n*-place connective ">" with a rigidly two-place combination ".~". This replaces a simple, streamlined, and aesthetically pleasing function with a clunkier and less graceful one that (as we have seen in Section 2, above) makes the move from (T') to (O) clumsier. But perhaps there will be compensations.

Similarly, let

(I^1) $Pw(x \vee y)$

be an *indifference* [or *I*-] *function* defined in terms of *P* that states that *w* prefers either *x* or *y*, i.e. either one is acceptable. Now were I to follow Savage's lead, I would define the indifference relation strictly in terms of (P) rather as

(I^2) $\sim Pw(x.\sim y).\sim Pw(y.\sim x).$

But Savage's definition (2. (I)), and (I^2) even more precisely, express, if anything specific to preference, something closer to revulsion for both *x* and *y* rather than true indifference between them. It is the difference between asserting that neither is preferred to the other – think of this as

Stoic indifference, and asserting that either one is fine – think of this as *Epicurean indifference*. If either one is fine, then either one is preferable – or both are equally so, in which case it is false that neither is preferred to the other. Given a pairwise comparison between Stoic and Epicurean indifference, I adopt the latter and reject the former, on the grounds that Epicurean indifference expresses a healthier outlook on life. What this comes to is a strict preference for the intuitive and linguistic benefits of capturing the strict preference and indifference relations in terms of the Boolean connectives more generally, over aspiration to Savage's stoically austere aesthetic standards. But I explore some further considerations that justify this preference below.

Nevertheless, (P) and (I^1) each can be defined in terms of the other to the following extent:

(P) $Pw(x.\sim y) =_{df.} Pw[(x \vee y).(\sim x \vee \sim y).(x \vee \sim y)]$

(I^1) $Pw(x \vee y) =_{df.} Pw[(x.\sim y) \vee (y.\sim x) \vee (x.y)]$

Both (P) and (I^1) embed w's satisfaction of conditions (2.a)-(2.b) above for being a conscious and intentional chooser with a genuine preference.

Within a single occurrence of P, the rules of inference that apply to sentential propositions and sentential functions apply also to the individual variables between the outermost brackets it contains. So (P) can be viewed as a result of using the same translation rules that lead from (3.1) to (3. 2), but on subsentential constituents rather than sentential propositions, thus:

(6) $Pw\{[(x \vee y).\sim(x.y)].\sim y\}$.

(6) says that w prefers either x or y but not both, and not y. (6) then can be transformed using the canonical rules of inference for sentential logic as follows:

(7) $Pw\{[(x \vee y).(\sim x \vee \sim y)].\sim y\}$

(8) $Pw[(x .(\sim x \vee \sim y)]$

(9) $Pw(x.\sim y)$.

Similarly,

(10) $Pw(x.\sim y)$

(11) $Pw\sim(\sim x \vee y)$

(12) $Pw \sim(x \rightarrow y)$

are all equivalent, and

(13) $Pw[(r.s).\sim(t \vee u)]$

is – like (6) – a legitimate substitution instance of (P). Call this the *subsentential application* of rules of logical inference.

This application may be useful in case a chooser is presented with a pair of alternatives each of which has a complex substructure of the sort explored in (3. 7) – (3. 10), above. For example, suppose Victor prefers beans or veggies indifferently over beans or rice indifferently, i.e.

(14) $Ps[(a \vee b).\sim(a \vee c)]$.

Then since (14) is equivalent to

(15) $Ps[(a \vee b).\sim a.\sim c]$,

which reduces to

(16) *Ps(b)*,

we can conclude – and encourage Victor to reason his own way to the conclusion – that what he really prefers is veggies *simpliciter*.

Variables *a*, *b* and *c* contained within brackets in (14) – (16) denote subsentential constituent objects of a single intentional attitude, namely Victor's singular preference *P*. So the familiar opacity restrictions do not apply to inferences and syntactical transformations among *a*, *b* and *c*. However, we have already seen that *P* is an intentional operator, so those restrictions do apply to rules of inference across multiple occurrences of such operators. Here we avoid opacity problems by stipulating similarly that given multiple occurrences of *P* or its substitution instances, the rules of inference governing sentential propositions and sentential functions apply conventionally to the relations among them. So, for example, from

(17) *Pw(x.~y)* → *Pw(x.~z)*

and

(18) *Pw(x.~y)*,

it is legitimate to infer

(19) *Pw(x.~z)*.

Excluded by this stipulation, however, would be any subsentential application of such rules to individual variables across multiple occurrences of *P* or its instances. So, for example, from

(20) *Pw(x.~y).Pw[(x→~z)* v *y]*

we could not infer

(21) *Pw(x.~z)*,

for this would be to apply the rules of inference subsententially to individual variables across two occurrences of *P* in a way that trespassed the outermost-bracketed boundaries between them. So it would be to violate what I shall call the *No-Trespass Rule*. We shall see shortly that the No-Trespass Rule has important implications for the reformulation of (T).

The No-Trespass Rule also enables us to distinguish between not having a preference for *x* over *y*, i.e.

(22) *~Pw(x.~y)*,

and positively preferring not to select *x* over *y*, i.e.

(23) *Pw ~(x.~y)*,

which would be equivalent to being indifferent between not selecting *x* and selecting *y*, i.e.

(24) *Pw(~x* v *y)*.

In (22) the tilda stands outside the *P*-function, and so modifies a sentential proposition. In (23), by contrast, it stands outside the outermost brackets of the P-function and so modifies only the subsentential constituent complex intentional object that lies within them. In (24) the scope of the tilda is even smaller: it modifies only the variable *x* to the right. But both (23) and (24) demonstrate how this Boolean connective might modify subsentential constituent intentional objects of a preference, not just the sentential assertion of that preference itself.

Putting this much to work, we can now reduce Luce and Raiffa's gloss on Savage's indifference relation (I) to what I call pure Epicurean indifference (I^1). Where x and y are acts, Luce and Raiffa define indifference between x and y (á la Savage) as: x is preferred to or indifferent to y and y is preferred to or indifferent to x, i.e.

(I^3) $x \geq y$ and $y \geq x$.[10]

Luce and Raiffa's substitution of "≥" for Savage's interpretation of "≤" (Section 2, above) sacrifices the latter's elegant yet cumbersome formulation to the demands of greater simplicity. In my suggested quasi-quantificational notation, (I^3) would run as follows:

(I^4) $Pw[(x.\sim y) \vee (x \vee y)].Pw[(y.\sim x) \vee (y \vee x)]$.

But like (6) above, (I^4) can be transformed as follows according to the familiar rules of inference for sentential logic:

(25) $Pw\sim[\sim(x.\sim y) . \sim(x \vee y)].Pw\sim[\sim(y.\sim x) . \sim(y \vee x)]$

(26) $Pw\sim[(x \rightarrow y) . \sim x.\sim y].Pw\sim[(y \rightarrow x) . \sim y.\sim x]$

(27) $Pw\sim[(\sim y \rightarrow \sim x) .\sim x.\sim y].Pw\sim[(\sim x \rightarrow \sim y) .\sim y.\sim x]$

(28) $Pw\sim[\sim x.\sim y] .Pw\sim[\sim x.\sim y]$

(29) $Pw[x \vee y] . Pw[x \vee y]$

(I^1) $Pw(x \vee y)$.

Thus Luce and Raiffa's Savagean definition of indifference is reducible to my notion of Epicurean indifference. With this notation we lose yet more aesthetically, as we move even further away from Savage's original conception. But we gain something more important, namely commensurability with other relations susceptible to sentential and subsentential analysis via the standard Boolean connectives.

Finally, we can define a *weak preference* [or *R*-] *function* in the conventional way, in terms of strict preference and Epicurean indifference, such that given two alternatives x and y,

(R^E) $Pw[(x.\sim y) \vee (x \vee y)]$

Section 6.2.1 below suggests a way of constructing the quick truth table that shows that (R^E) is a tautology. (R^E) states that w prefers either one alternative to the other strictly, or else either one is fine. Weak preference on this rendering is that special case of Epicurean indifference in which either strictly preferring one alternative or finding either one fine is itself fine. Then following Edward McClennen's results with conventional notation, (P), (I^1), and (R^E) can be particularized to gambles g_1 and g_2 as follows:

(30) $Pw(x.\sim y) \equiv Pw(g_x.\sim g_y)$

(31) $Pw(x \vee y) \equiv Pw\ (g_x \vee g_y)$

(32) $Pw\ (x.\sim y) \rightarrow Pw[(g_x.\sim g_y) \vee (g_x \vee g_y)]$

(33) $Pw\ (x \vee y) \rightarrow Pw\ [(g_x.\sim g_y) \vee (g_x \vee g_y)]$.

Clearly McClennen's Particularization Principle has equal application to actions, states, events, compound lotteries, plans, prospects, and discrete objects.[11] By contrast, weak preference defined using Stoic indifference,

(R^S) $Pw\{(x.\sim y) \vee [\sim Pw(x.\sim y).\sim Pw(y.\sim x)]\}$,

is a contingent truth which states that *w* either prefers one alternative to the other strictly, or else prefers neither to the other.

That the variable term calculus makes (R^E) tautological is in its favor. For it makes having a weak preference between two alternatives a conceptual truth about what it means to have a preference at all, namely that either one alternative is clearly superior, or else either alternative is acceptable. By contrast, if neither alternative is sufficiently attractive to be acceptable, one cannot be said to have a genuine preference in the first place. This notion of neither alternative being acceptable was the basic idea Savage's original formulation of the indifference relation attempted to capture. But it is too strong to enter into a definition of weak preference. Yet this is the condition that the second disjunct of (R^S) expresses. (R^S) states that one either has a strict preference or else no preference at all, but instead merely generalized revulsion for *x* and *y*. That is far too strong. (R^S) cannot be a *definition* of weak preference because it is not always true wherever a preference exists.

So the notation of the suggested variable term calculus enables us to detect a logical and conceptual conflict between Savage's original, Stoic conception of the indifference relation (R^S), and Luce and Raiffa's Epicurean gloss on it (R^E). This conflict becomes salient when we try to plug Savage's conception into a concept of weak preference, conventionally defined as strict-preference-or-indifference between two alternatives. We see that (R^S) therefore attempts to define weak preference as either strict preference or no preference. But no preference is not an option for a chooser who claims to have a preference of some sort or other. The two concepts contradict each other, as we can see:

(34) $Pw[(x.\sim y).[\sim Pw(x.\sim y).\sim Pw(y.\sim x)]$.

Of course someone who chooses to use this notation to build an axiomatic system is free either to follow Savage and use (R^S) to define weak preference; or else to follow Luce & Raiffa – and my proposed variable term calculus – and use (R^E). But in this section I have tried to offer several good reasons for strictly preferring (R^E). Epicurean indifference is healthier both psychologically and logically.

6. Indifference, Indecision, and Equivalence

6.1. Kaplan on Rational Indecision

By way of contrast with my analysis of indifference, consider Mark Kaplan's unorthodox Bayesian approach. Kaplan aims to argue against what he characterizes as "the sin of false precision," [12] i.e. the Ramseyan practice of assigning monetary values to most states of affairs constitutive of gambles, i.e. of assigning determinate degrees of confidence in hypotheses for the truth of which we have either insufficient evidence or no evidence. He believes that "most of the hypotheses we have occasion to investigate in ordinary life are ones to which we are not warranted in assigning a determinate degree of confidence."[13]

6.1.1. Preference and Indifference

> Kaplan defines something A as a *well-mannered state of affairs* just in case for some set of mutually exclusive and jointly exhaustive hypotheses $\{P_1, \ldots, P_n\}$, and some set of real numbers $\{a_1, \ldots, a_n\}$, A = ($\$a_1$ if P_1, . . ., $\$a_n$ if P_n). If all a_i are equal to the same sum a, A = $\$a$ [DTP 4].

Then given well-mannered states of affairs A and B, either you prefer A to B, or you prefer B to A, or you prefer neither A to B nor B to A, in which case you are either indifferent between A and B, or undecided between A and B. Kaplan explains the difference between indifference and indecision as follows:

> When you are indifferent between A and B, your failure to prefer one to the other is born of a determination that they are equally preferable. When you are undecided, your failure to prefer one to the other is born of no such determination [DTP 5].

However, according to the conclusion of Section 5 above, failing to prefer one to the other cannot arise out of a determination that they are equally preferable, because the two are mutually contradictory. Failure to prefer A to B and B to A is, substituting x for A and y for B, Savage's Stoic conception of indifference:

(I^2) $\sim Pw(x.\sim y).\sim Pw(y.\sim x)$

Given only these two alternatives, if you fail to prefer A to B and fail to prefer B to A, then you at least fail to prefer A and fail to prefer B. So from (I^2) we can infer

(1) $\sim Pw(x).\sim Pw(y)$.

But if A and B are equally preferable, then you prefer A and you also prefer B:

(2) $Pw(x).Pw(y)$.

But then failing to prefer either A or B cannot, on pain of contradiction, arise out of A and B being equally preferable. The Stoic conception of indifference precludes this.

Kaplan might respond by distinguishing, as Mill failed to do, between A and B as equally preferable, and A and B as equal in preferability. To say of something that it is preferable is to make a positive value judgment about it – in much the same way we do when we judge something to be desirable. So to say of two things that they are *equally preferable* is to make the same positive value judgment about both of them. It is to express a favorable intentional attitude toward them. Describe this as "preferableness." By contrast, to say of two or more things that they are *equal in preferability* is to make a value-neutral modal judgment about their capacity or potential *to be* preferred, or preferable. It is to make an extensional metaphysical claim about their properties. A and B can be equal in preferability consistently with neither being preferred to the other; this is just to say that their metaphysical capacity or potential to be the object of this favorable intentional attitude is zero. Thus preferableness is distinct from preferability.

However, when using the concept of preference in a theory of decision, we cannot mean to be using the metaphysical concept of preferability, on pain of irrelevance. That is, when asked which, between A and B, we strictly prefer, we cannot mean to evaluate them on the basis of which has a greater *capacity or potential* to be preferred. Who cares? We need to decide which,

between A and B, we actually prefer. Similarly, when we state our indifference between A and B, we cannot mean to evaluate both as equal in capacity or potential for being preferred, or neither as superior to the other in this capacity, because this extensional metaphysical assessment has nothing to do with the choice before us. Nor can we use the intentional concept of preferableness to dissect strict preference on the one hand, while using the metaphysical concept of preferability to dissect indifference on the other, on pain of inconsistency. Therefore, although it is true that, as stated above, A and B can be equal in preferability consistently with neither being preferred to the other, this cannot be a viable analysis of decision-theoretic indifference.

Another possibility would be to explain Stoic indifference between A and B as the case in which you fail to prefer A to B and fail to prefer B to A, because you regard A and B as either equally preferable or equally unpreferable. In this case, you fail to prefer A and fail to prefer B because either you fail to prefer A and fail to prefer B, or you prefer A and prefer B. This account of Stoic indifference between A and B is tautologically true. It says that when you do not prefer either A or B, this is because either you do not or you do prefer either A or B. The difficulty is that the very same account can be given of preferring A and B equally. In the proposed variable term calculus, the explanation would run this way:

(I^5) *[Pw~(x.~y). Pw~(y.~x)]* → *{[Pw~(x.~y). Pw~(y.~x)]* v *[Pw(x.~y).Pw(y.~x)]}*

Clearly, (I^5) is a tautology. So is the following:

(3) *[Pw(x.~y).Pw(y.~x)]* → *{[Pw~(x.~y). Pw~(y.~x)]* v *[Pw(x.~y).Pw(y.~x)]}.*

Both (I^5) and (3) have the same tautological structure, namely that if P then P or Q. The consequent does not succeed in explaining anything.

On Kaplan's account, failing to prefer A to B and B to A is a broader notion of which both indifference and indecision are special cases. He says, "mere failure of preference ... does not constitute indifference – does not constitute something that can reasonably thought to conform to transitivity."[14] I agree with this, but my reasons are different than Kaplan's. On my Epicurean account, indifference is inconsistent with failing to prefer – i.e. what I above called generalized revulsion; and Kaplan's use of the term "indecision" covers two kinds of failure to decide that need to be further distinguished: true indecision and decisional incapacity.

6.1.2. Indecision and Decisional Incapacity

True indecision is the special case of failing to decide in which I am ambivalent between A and B. But there are other cases in which I may not merely fail to decide, but be in addition unable to decide: because I lack either the ability, or the opportunity, or the permission, or the authority, or the resources, or some conjunction thereof. For example, I am unable, in this sense, to decide whether NASA should postpone a rocket launch to Mars until next year, because I lack all of the above. For these cases, the standard terminology is insufficiently fine-grained: inability to decide in this sense is not adequately analyzed in terms of failure to *prefer* either option, since I can have preferences about matters over which I lack decision-making power. So instead let us analyze inability to decide in terms of failure to *choose* A over B or B over A. I fail to make that

choice when, because of any of the factors just listed, I lack the power to make that choice. I shall describe this kind of failure to decide as *decisional incapacity*. Kaplan is correct in so far as he means to deny that decisional incapacity necessarily satisfies transitivity.

However, I also may fail to decide because I lack the certitude – i.e. because I am truly *undecided*. When I am undecided it is not because I lack the ability, opportunity, permission, authority, or resources needed to decide. It is because I am simply ambivalent between the alternatives. Whereas decisional incapacity is a metaphysical state, indecision – like preference and indifference – is a psychological one. When I am ambivalent between two alternatives, I may similarly fail to decide – or choose – between them. But not necessarily because I fail to *prefer* A to B and B to A. If indecision between A and B is true ambivalence between A and B, then indecision no more implies failure to prefer either to the other than it implies actively preferring either to the other. Indecision is not a way of failing to prefer at all, because ambivalence does not necessarily imply failure to prefer. Rather, indecision is an entirely separate phenomenon the logic of which needs to be developed independently.

Hence the Stoic notion of failing to prefer A to B and B to A is not useful as an interpretation of indifference because, *contra* Kaplan, it cannot be cashed out in terms of equal preferability. It is not useful as an interpretation of decisional incapacity because one can have quite determinate preferences about alternatives between which one lacks the resources to decide. It is not useful as an interpretation of indecision because it does not necessarily hold in the quintessential case of indecision, namely ambivalence. Finally the Stoic notion therefore does not necessarily hold in all cases of inability to decide.

Now Kaplan constructs his argument against false precision by imposing the following rational constraints, among others, on choices among well-mannered states of affairs:

(1) *Ordering*: Where A, B and C are any well-mannered states of affairs between no pair of which you are undecided,

(i) you do not prefer A to A [i.e. irreflexivity];

(ii) if you do not prefer A to B and you do not prefer B to C, then you do not prefer A to C [i.e. transitivity for weak preference] [DTP 5].

(2) *Dominance*: Where A is ($*a* if P_1, . . ., $\$a_n$ if P_n) and B is ($\$b_1$ if P_1, . . ., $\$b_n$ if P_n),

(i) if $a_i > b_i$ for every *i*, then you prefer A to B; and

(ii) if $a_i = b_i$ for every *i*, then you are indifferent between A and B [DTP 7].

(3) *Confidence*: For any hypotheses P and Q, you are more confident that P than you are that Q iff you prefer ($1 if P, $0 if ~P) to ($1 if Q, $0 if ~Q) [DTP 8].

(4) *Decomposition*: A sequence of well-mannered states of affairs ϕ is a *decomposition of* a well-mannered state of affairs A (i.e. A *is composable from* ϕ) just in case ϕ is finite and it is logically impossible that the realization of ϕ will effect … a different net change in your fortune than the realization of A will. You *place a monetary value of* $*a* on A just if you are indifferent between $*a* and A. Then if

(i) A is a well-mannered state of affairs;

(ii) ϕ is a decomposition of A; and

(iii) you place a monetary value on A and on each of the terms of ϕ;

then the value you place on A is equal to the sum of the values you place on the terms of ϕ [DTP 10].

(5) *Modest Connectedness*: Your preferences are characterized by a non-empty set V of assignments of monetary values to all well-mannered states of affairs such that each assignment satisfies Ordering, Dominance and Decomposition, and such that

(i) you are indifferent between A and B just in case every member of V assigns A the same monetary value as it assigns B; and

(ii) you prefer A to B just in case no member of V assigns B a greater monetary value than it assigns A and some member of V assigns A a greater value than it assigns B [DTP 13].

From (3) and (5), Kaplan derives

(6) *Modest Probabilism*: Any assignment of monetary values to all well-mannered states of affairs that satisfies Ordering, Dominance and Decomposition assigns a real number to *con*(P) for every hypothesis P such that, for any hypotheses P and Q in a non-empty set W of con-assignments,

(i) *con*(P) ≥ 0;

(ii) if P is a tautology, then *con*(P) = 1;

(iii) if P and Q are mutually exclusive, then *con*(P v Q) = *con*(P) + *con*(Q);

(iv) you are just as confident that P as you are that Q just in case, on every member of W, *con*(P) = *con*(Q); and

(v) you are more confident that P than that Q just in case, on no member of W, *con*(Q) > *con*(P) and, on at least one member of W, *con*(P) > *con*(Q) [DTP 16, 21].[15]

These constraints are brought to bear on the following decision problem:

> Suppose *h* = the hypothesis that the incumbent will win re-election, and ~*h* = the hypothesis that she will lose; and that you know only that the incumbent is ahead of her only opponent. By contrast, *g* = the hypothesis that the ball drawn from an urn which you know contains exactly 50 black balls and 50 white ones is black; and ~*g* = the hypothesis that the ball drawn is white. So you are more confident that *h* than that ~*h*, and just as confident that *g* as that ~*g*. Next suppose you are offered a choice between (i) a ticket that yields you $1 if *h* and 0$ if ~*h*; and (ii) a coupon that yields you $1 if *g* and $0 if ~*g*; that dollar values are the only ones at issue; and that the worth to you of each dollar remains the same regardless of how many of them you have. Which should you choose, the ticket or the coupon? [DTP 1-2]

(i) seems attractive because you have reason to be more confident that *h* than that ~*h*. But your confidence is based on very little information. (ii) seems attractive because even though *g* and ~*g* are equally likely, you have full information about both possibilities. Kaplan argues convincingly that you should choose the ticket.

But then he poses the following dilemma: suppose we alter the choice so that instead of the coupon, option (ii) offers $0.75. So now you must choose between (i) $1 if *h*, $0 if ~*h*; and (ii') $0.75. Since by hypothesis you are more confident that *h* than that ~*h*, constraints (3), (5) and (6) above require you to exclude assignments of monetary values to well-mannered states of affairs in W of less than $0.50 to ($1 if *h*, $0 if ~*h*), and to include such assignments greater than $0.50. But that means including some such assignments greater than $0.75, and some that are less. Because you cannot assign a precise degree of confidence in the likelihood of *h* as opposed to ~*h*, no decision between (i) and (ii') can be made.

Kaplan interprets this dilemma as a case of indecision because in it you fail to prefer – or better, *choose* – A over B or B over A. However, the distinction on which I have insisted above suggests that this case is better understood as one of decisional incapacity; i.e. in which, given the stipulated constraints, you are instead unable to decide – i.e. you lack the ability, opportunity, and resources you need to decide – between (i) and (ii'). Of a piece with this case would be any similar one in which one lacks sufficient information about A, or in which, for example, A is a bet on an hypothesis for which one has little evidence, no evidence or for which one can have no evidence. True indecision, by contrast, occurs when the obstacle to choice is not a lack of evidence but rather ambivalence.

We might expect the logic of indecision to be different from the logic of decisional incapacity. Since indecision, on my account, does not imply decisional incapacity, evaluating its satisfaction of rational requirements on choice would be appropriate. By contrast, the greater the decisional incapacity, the more those requirements must be relaxed. Kaplan's argument against false precision is of this kind. It is in the tradition that relaxes the rationality requirements for fully informed choice to those of risky choice; then relaxes these to the less stringent requirements of choice under uncertainty. Kaplan's thesis advocates a further relaxation of those requirements that adapt to the circumstance in which one's evidence is of insufficient quality to warrant a determinate assignment of confidence to one or more of the hypotheses on which the outcome of a gamble depends. He concludes that Connectivity must be abandoned under these conditions. Another option would be to abandon as a general practice the assignment of determinate degrees of confidence over gambles. But that would be an unpopular alternative.[16]

6.2. Indifference and Equivalence in the Jeffrey-Bolker Representation Theorem

Fixing the appropriate rendering and interpretation of the indifference relation, and distinguishing between it and other concepts with which it might be conflated, is important because of the problem encountered in Volume I, Chapter IV.2.2. in Ramsey's exposition of the axioms by which a subjective measure of cardinal utility is constructed. We saw there that Ramsey's move from an indifference relation in (A1) to an equivalence relation in (A2) was problematic, because the two relations do not seem to have the same properties and Ramsey's own exposition offered no defense of the move. That move was nevertheless significant because it enabled Ramsey to exchange an intensional relationship of necessarily limited quantitative

scope for an extensional one with much greater quantitative range and flexibility. Without it, construction of the subjective probability measure on which Kaplan's unorthodox Bayesian approach relies would be much more cumbersome (though not necessarily impossible). The Jeffrey-Bolker representation theorem does, however, actually defend a more streamlined version of Ramsey's move. But in order to evaluate it, it will be useful to introduce a bit more apparatus, familiar from classical logic but suitably modified for subsentential constituents.

6.2.1 Occasional Truth Tables for Subsentential Constituents

In conventional logic, a truth table enables us to assess whether two (or more) complex statements, structured by the Boolean connectives, are consistent, equivalent, tautological, self-contradictory, or contingent, by consistently assigning the values **T** or **F** to each sentence letter that each such complex statement contains. The reason we can do this is that classical logic is an extensional symbolic language whose connectives relate sentences or propositions that are reasonably assumed to be true or false independently of the complex statements in which they are embedded.

The variable term calculus I am proposing here requires modification of this assumption. On the one hand we can, analogously, assign truth-values to terms and variables within a single intentional attitude in order to fix the truth or falsity of preferences the agent may have at a particular moment. A *true preference Ps(a.~b)* would be one that assigns **T** to *a* and **F** to *b*. A *false preference Ps(a.~b)* would be one that assigns **F** either to *a* or to *~b*. Instead of speaking of true or false sentences or propositions as denoting or failing to denote a state of affairs, we would speak of true or false preferences as denoting or failing to denote a particular intentional state of the agent. On the other hand, because we are working within the constraints of an intensional language, there can be no guarantee that the truth-value of a variable or term that occurs within the scope of one intentional attitude (i.e. such that its occurrences are enclosed within the outermost brackets governed by *Pw* or its instantiations) will be the same as its truth-value within the scope of a different one. So, for example, it is possible that in the following statement

(1) *Ps(a.~b).Ps(a.~c).Ps(c.~a)*,

the truth assignments in the third conjunct might be the reverse of what they are in the first two. Because this is always a possibility, the usefulness of truth tables for intentional attitudes such as preferences – and indeed for the variable term calculus more generally – is limited. Truth tables for subsentential constituents are reliable only on those occasions in which the truth assignments to variables or terms are consistent over the range of intentional attitudes related by the Boolean connectives within a complex statement such as (1). Hence the description of these truth tables as *occasional*.

However, only under this restriction are the criteria of horizontal and vertical consistency fully satisfied. Since, as we have seen in Chapter II, satisfaction of these two criteria are necessary conditions of unified agency, presupposing them in this ideal scenario does not seem too much of a stretch (even though in reality, as I argue in Part II below, we often fail to satisfy these

conditions). This presupposition, in turn, enables us to compare the truth-values of two different intentional attitudes related by a Boolean connective, i.e. in the case in which each intentional attitude functions as would a sentence letter in sentential logic; and on the basis of these to assign a truth-value to the complex statement in which such connectives occur.

Following is a simple statement of second-order preference that would be an unobjectionable candidate for truth-functional analysis:

(2) $Pw\{x.\sim[Pw(x \vee y)]\}$

(2) says that an agent prefers x to being Epicureanly indifferent between x and y. Is (2) a consistent preference? Using Quine's method of establishing consistency and validity,[17] the following occasional truth table

Pw	*{x*	*.*	*~[Pw*	*(x*	*v*	*y)]}*
	T			**T**		**T**
					T	
			F			
		F				
F						

demonstrates that (2) is not a consistent preference: It does not make sense to say that one prefers some option on the one hand, but that that same option would not be fine, i.e. perfectly acceptable, if offered in a pairwise comparison of alternatives on the other.

A more complex statement for which truth-functional analysis yields interesting results – under the presupposition that horizontal and vertical consistency are satisfied – is the one I made in Volume I, Chapter III.1, that the transitivity and acyclicity axioms are logically equivalent. In the proposed variable term calculus, that statement would look like this:

(3) $\{[Pw(x.\sim y).Pw(y.\sim z)] \rightarrow Pw(x.\sim z)\} \equiv \{[Pw(x.\sim y).Pw(y.\sim z)] \rightarrow \sim Pw(z.\sim x)\}$

I discuss the implications of the left-hand statement in the above biconditional at greater length in Section 7, below. Whether (3) is true or not would turn on whether it was possible to assign truth-values to (3) that made one side of the biconditional false and the other side true. If so, (3) is false; if not, true. The truth-functional analysis that tests the validity of (3) would look this way (I break it into two tables, one for each side of the biconditional, for ease of reading).

Let us first try to make the left-hand side of the biconditional false:

{[Pw	*(x*	*.*	*~y)*	*.*	*Pw*	*(y*	*.*	*~z)]*	*→*	*Pw*	*(x*	*.*	*~z)}*
								F					**F**
							F					**F**	
					F					**F**			

				F									
									T				

We cannot. Since the left-hand side of the biconditional is **T** under all assignments, let us now try to make the right-hand side **F**, maintaining the same truth-value assignments as for the left-hand side, under the assumption of horizontal and vertical consistency:

{[Pw	*(x*	.	*~y)*	.	*Pw*	*(y*	.	*~z)]*	→	*~Pw*	*(z*	.	*~x)]}*
	F							**F**			**T**		**T**
							F					**T**	
				F						**F**			
									T				

It appears that the right-hand side of the biconditional also turns out **T** under all assignments. Hence so does the biconditional (3) itself. The transitivity and acyclicity axioms are logically equivalent. (2) and (3) demonstrate how occasional truth tables for subsentential constituents might work. With this additional apparatus let us now turn to the Jeffrey-Bolker representation theorem.

6.2.2. Is Indifference an Equivalence Relation?

Jeffrey-Bolker decision theory solves Ramsey's problem, of how to move from non-quantitative conditions on preference rankings to quantitative functions that represent those rankings cardinally and probabilistically, using the following reasoning.[18] Begin with a primitive notion of weak preference A≥B, such that A is preferred to or indifferent to B. Interpret this as meaning that A is at least as high as B in the agent's preference ordering. Assume that the weak preference relation "≥" is transitive and connected, such that

(1) If A≥B and B≥C then A≥C, and

(2) either A≥B or B≥A or both

respectively. Then use "≥" to define preference and indifference, as follows:

(3) Preference =df. A>B iff A≥B but not B≥A.

Show that preference satisfies Irreflexivity (not A>A), Asymmetry (if A>B then not B>A), and Transitivity (if A>B and B>C then A>C).

(4) Indifference =df. A ≈ B iff A≥B and B≥A

(This is also Luce and Raiffa's definition of indifference.) Show that indifference thus defined is an equivalence relation: it satisfies Symmetry (if A≈B then B≈A), Reflexivity (A≈A), and Transitivity (if A≈B and B≈C then A≈C).

Let us now distinguish between the intuitive notion of indifference that enters into the primitive weak preference relation, and the fully defined indifference relation as spelled out in (4). The Jeffrey-Bolker improvement on Ramsey's argument is to defend the claim that

indifference is an equivalence relation by arguing that the fully defined indifference relation satisfies these three criteria.

Let us grant that any viable notion of indifference must satisfy Symmetry and Reflexivity. But in Volume I, Chapter IV.2.2., in commenting on Ramsey's axiom (A3'), I offered some reasons to doubt whether the indifference relation always satisfied Transitivity; and intuitively, it is hard to see what is irrational about my indifference between cherries and apples and between apples and peaches, but strong preference for peaches over cherries, even if my pairwise comparisons adhere to a unidimensional criterion such as flavor. The reason for this is that the notion of indifference in play in these three pairwise comparisons is an intensional one. So I wish to press hard on the Jeffrey-Bolker thesis that indifference satisfies transitivity; and then to question what this implies even if it does.

First, if indifference is to be defined in terms of weak preference as an undefined primitive relation in (4), then how is the intuitive notion of indifference in the weak preference relation itself to be interpreted subsententially? There are two possibilities. One way would be to interpret it as Savage's Stoic indifference, i.e.

(5) $\sim Pw(x.\sim y).\sim Pw(y.\sim x)$.

In this case weak preference would look this way:

(6) $A \geq B =_{df.} Pw(x.\sim y) \vee [\sim Pw(x.\sim y).\sim Pw(y.\sim x)]$,

and the fully defined indifference relation in (4) like this:

(7) $A \geq B$ and $B \geq A =_{df.} \{Pw(x.\sim y) \vee [\sim Pw(x.\sim y).\sim Pw(y.\sim x)]\} .$
$\{[Pw(y.\sim x) \vee [\sim Pw(y.\sim x).\sim Pw(x.\sim y)]]\}.$

Another way to understand subsententially the intuitive notion of indifference in the weak preference relation would be in terms of my concept of Epicurean indifference, i.e.

(8) $Pw(x \vee y)$,

in which case weak preference would be rendered this way:

(9) $A \geq B =_{df.} Pw(x.\sim y) \vee Pw(x \vee y)$,

and indifference strictly speaking, i.e. as in (4), like this:

(10) $A \geq B$ and $B \geq A =_{df.} [Pw(x.\sim y) \vee Pw(x \vee y)].[Pw(y.\sim x) \vee Pw(y \vee x)]$.

Looking now at (1), above, it seems clear that a weak preference ordering is transitive only if the intuitive notions of preference and indifference that define it are. Let us grant the unidimensional transitivity of preference. What about indifference? Is the intuitive notion of indifference itself transitive in all cases, on either the Stoic or the Epicurean interpretation? This is the first question. A second is whether either interpretation of the intuitive notion of indifference makes the fully defined Jeffrey-Bolker indifference relation in (4) transitive in all cases. The last will be what this implies for the thesis that indifference is an equivalence relation.

Keeping in mind the restrictive presupposition of horizontal and vertical consistency mentioned in 6.2.1, we can call on an occasional truth table to suggest answers to these questions.

1. Is the intuitive notion of indifference itself transitive in all cases?

Take first the Stoic interpretation. (5) above can be plugged into a transitivity rule as follows:

(11) *{[~Pw(x.~y).~Pw(y.~x)].[~Pw(y.~z).~Pw(z.~y)]} → [~Pw(x.~z).~Pw(z.~x)]*

From the following truth table, it appears that (11) is true under all assignments to x, y, and z:

{[~Pw(	*x*	*.*	*~y)*	*.*	*~Pw(*	*y*	*.*	*~x)]*	*.*	*[~Pw(*	*y*	*.*	*~z)*	*.*	*~Pw(*	*z*	*.*	*~y)]}*
	F		**F**			**T**		**T**			**T**		**F**			**T**		**F**
		F					**T**					**F**					**F**	
T					**F**					**T**					**T**			
				F										**T**				
									F									

→	*[~Pw(*	*x*	*.*	*~z)*	*.*	*~Pw(*	*z*	*.*	*~x)]*
							T		**T**
								T	
						F			
					F				
T									

(11), then, is a tautology. Hence on the Stoic interpretation, the intuitive notion of indifference satisfies Transitivity in all cases, in addition to Symmetry and Irreflexivity. There are two ways of reading this result. One is as a vindication of the Jeffrey-Bolker thesis. A second is as further evidence of the Stoic interpretation's inadequacy to capture the character of indifference as an intentional attitude. For it fails to accommodate the seemingly unobjectionable case of being indifferent between cherries and apples and between apples and peaches, but of having a strong preference for peaches over cherries. On the Stoic interpretation, this is simply irrational. If the Stoic interpretation is mistaken, then the second reading of this result is preferable.

By contrast, (8) above – the Epicurean interpretation of the intuitive notion of indifference, when plugged into a transitivity rule, looks this way:

(12) *[Pw(x* v *y) . Pw(y* v *z)] → Pw(x* v *z)*

As we can see below, (12) fails transitivity in case *x*=**F**, *y*=**T**, *z*=**T**:

[Pw(	*x*	v	*y)*	*.*	*Pw(*	*y*	v	*z)]*	→	*Pw(*	*x*	v	*z)*
	F		**T**			**T**		**F**			**F**		**F**
		T					**T**					**F**	
T					**T**					**F**			
				T									
									F				

This is to be applauded, because it leaves room for those unobjectionable cases the Stoic interpretation excludes. But if there exists even one such case in which intuitive Epicurean indifference is intransitive, then either it would seem not to be a good candidate for an indifference relation, fully defined on the concept of weak preference as in (4), that purports to be an equivalence relation; or else it indicates that on the more plausible interpretation of intuitive indifference, indifference is not an equivalence relation. Either way, the Epicurean interpretation of intuitive indifference makes more stringent demands, of an intensional nature, on a representation theorem that formulates nonquantitative conditions on preference as quantitative preference and probability functions.

2. Does either interpretation of the intuitive notion of indifference make the fully defined Jeffrey-Bolker indifference relation in (4) transitive in all cases?

It transpires that a truth-functional analysis shows that both the Stoic and the Epicurean interpretations of the intuitive notion of indifference makes the fully defined indifference relation in (4) fully transitive in all cases, although the truth table is too cumbersome to reproduce here. On either interpretation of the intuitive notion, (4) satisfies all three of the conditions the Jeffrey-Bolker thesis requires. The contingent transitivity – i.e. intensionality – of the intuitive notion under the Epicurean interpretation is obscured when it is brought into the more complex definition of the indifference relation in (4), in the same manner in which the contingency of (P.~Q) may be obscured when folded into a compound tautological sentence like [(P.~Q) → (P.~Q) v ~(P.~Q)], or the inconsistency of (P.~P) may be when buried in a valid compound sentence like [(P.~P) → (P v Q)]. In none of these cases does the compound sentence conceal the logical import of the subsentential constituent. So this result opens the door to raising further questions about the relation between the intuitive, intensional notion of indifference that enters into weak preference, and the more complex version of indifference that is stipulated to build upon it. In particular, it raises the question whether a compound sentence *can be* fully extensional if it includes an intensional subsentential constituent: Could the arithmetical sentence

(13) 2 + 2 = 4

be fully extensional if, for the first conjunct, we substituted the constituent sentence, "Piper is indifferent between 2 and 1+1"? Is the resulting sentence,

(14) (Piper is indifferent between 2 and 1+1) + 2 = 4

extensional? I think not. But we need not resolve the question here. For present purposes it is enough to have shown how an occasional truth table for subsentential constituents can expose the intensionality of the intuitive notion, and to note that these questions, about the purported extensionality of the complex sentences that depend on it, can be raised.

3. Does this answer to Question 2 make indifference an equivalence relation?

That is, does the fact that on either interpretation of indifference, the fully defined indifference relation satisfies all three conditions – Symmetry, Irreflexivity, and Transitivity – suffice to identify indifference as an equivalence relation? In case you are not convinced by the foregoing considerations, the same counterarguments to Ramsey offered in Volume I, Chapter IV.2.2 also apply here, and militate against a positive answer to this question. To say that I am *indifferent* between two choice alternatives x and y is to say that x and y occupy the same position in my preference ranking; that either one will do. By contrast, to say that x and y are *equivalent* is to say that x is a necessary and sufficient condition for y. It is to say first that if x is a choice alternative then y is also one; and that if y is a choice alternative then x is also one. It is to say that x is a choice alternative if and only if y is. However, that two choice alternatives occupy the same position in my preference ranking neither implies nor suggests any such relations of logical necessity between them. So the answer to this question is no: Satisfaction of Symmetry, Irreflexivity, and Transitivity does not suffice to make indifference an equivalence relation.

4. Does this show the impossibility of moving from non-quantitative ordering conditions on preference rankings to quantitative functions that represent those rankings cardinally and probabilistically?

I do not see why it should, since we hold intentional attitudes toward quantitative functions. The von Neumann-Morgenstern method of constructing a cardinal utility measure (discussed in Volume I, Chapter IV.1.2) does not require the assumption that indifference is an equivalence relation. Why mightn't this method be adequately modified and appropriated into a representation theorem based on subjective probabilities? However, I merely throw out these suggestions without attempting to answer them.

7. Criteria for a Genuine Preference

Next I propose five normative criteria that selection behavior must satisfy in order to qualify as a genuine preference. Essentially these amount to formalizing (a) and (b) in Section 2 in the suggested variable term calculus. They avoid the criticism I mounted of orthodox normative decision-theoretic axiom systems in Section 1, i.e. that they restrict the scope of application of the system, while at the same time failing to eliminate cyclical preferences from the wider empirical realm of logical possibility within which the axiom system is nested as a special case. Like classical logic, and unlike orthodox decision-theoretic axiom systems, the five criteria that follow mirror the limits of logical possibility in empirical reality. They eliminate cyclical preferences as a logical possibility by definition of what a genuine preference is. On this definition, the logical impossibility of cyclical preference follows as a conceptual truth.

Let a genuine – i.e. a logically consistent – preference $Pw(x.\sim y)$ satisfy the following five criteria:

(Asy) $Pw(x.\sim y).\rightarrow\sim Pw(y.\sim x)$ (Asymmetry)

(Asy) implies that if, for example, Una prefers veggies to rice, then she does not prefer rice to veggies. Savage thinks that (Asy) is implied by the very meaning of preference, and I shall follow him.[19]

(Con) $Pw[(x.\sim y)\ \text{v}\ (x\ \text{v}\ y)]\ \text{v}\ Pw[(y.\sim x)\ \text{v}\ (y\ \text{v}\ x)]$ (Connectivity)

(Con) says that, given any set S of alternatives *x*, *y*, *z*, …, any two alternatives in the set are such that one is either strictly preferred, indifferent, or weakly preferred to the other; i.e. that each alternative in the set stands in a defined preference relation to each of the others.

(Irr) $\sim Pw(x.\sim x)$ (Irreflexivity)

(Irr) says that no one prefers an alternative to itself. I take this criterion, too, to be implied by the very meaning of preference. Note its formal similarity to the axiom of nonself-contradiction (4.II') and to Chapter II.4.1's (HC). (Irr) is, in fact, what nonself-contradiction – i.e. horizontal consistency – comes to for noncomparative preferences. Together with (Asy), it imposes analogous restrictions on pairwise comparisons, since if

(1) $Pw(x.\sim x)$,

then by substitution on (Asy),

(2) $Pw(x.\sim x) \rightarrow \sim Pw(x.\sim x)$,

which implies

(3) $Pw(x.\sim x).\sim Pw(x.\sim x)$,

in which case self-contradiction abounds (upon which more below, Section 11).

(T^3) $Pw(x.\sim y).Pw(y.\sim z). \rightarrow Pw(x.\sim z)$ (Transitivity)

Recall that we earlier made use of (T^3) in Section 6.2.2 above, where we showed its logical equivalence to the acyclicity axiom. (T^3) says that if, for example, Bertram prefers veggies to rice and rice to beans, then he prefers veggies to beans. (T^3) is the time-independent, logically consistent rule applied by a chooser who is able both to form and apply the concept of some one thing's ranking superiority consistently over a series of pairwise comparisons (condition (2. a) of being a conscious and intentional chooser), and also to remember the relation of the two alternatives she is presently ranking to the third she is not (condition (2. b)). For when she prefers *y* to *z* at t_2, she remembers having preferred *x* to *y* at t_1. That is, she remembers at t_2, while ranking *y* and *z*, that there is also an *x* such that she prefers *x* to *y*, *as* she is ranking that very same *y* over *z*.

This is what Kant would call "reproduction of the manifold in imagination." But it is also what lies behind Savage's observation that

> I find on contemplating the three alleged [cyclical] preferences side by side that at least one among them is not a preference at all, at any rate not any more.[20]

That is, a cyclical "preference" depends on a failure to properly conceptualize one's selection behavior as the expression of a genuine preference, and a consequent failure to remember all three alternatives simultaneously. It thus depends on a failure to satisfy conditions (2. a) and (2. b) of being a conscious and intentional chooser.

A few further words about (T^3). If we were to violate the No-Trespass Rule, the two sets of bracketed individual variables conjoined in (T^3)'s antecedent would be interpretable as containing a contradiction, and therefore (T^3) would be what we might call a *bad tautology*. But violating the No-Trespass Rule would open precisely the Pandora's box of problems about the intensional opacity of *P* the rule itself is designed to eliminate (more on this in Section 10, below). Furthermore, violating the rule would produce the kind of flatfooted interpretation of preference – such that *w* would be guilty of the intertemporal logical inconsistency of preferring *y* least at t_1 and most at t_2 *simpliciter* – that is foreclosed by my earlier formulation of (P) in Section 5, above.

By contrast, observation of the No-Trespass Rule for (P) solves these two problems. It embeds the two interconnected concepts of a conscious and intentional chooser and a genuine preference in such a way as to require that *w*'s preference ranking of *y* over *z* at t_2 be intertemporally logically consistent with her ranking of *x* over *y* at t_1, i.e. such that *Pw(x.~z)* is true *by implication*. This is part of what it means to describe (T^3) as a conceptual truth. *P*'s intensional opacity requires observation of the No-Trespass rule. But this has the felicitous side-effect of eliminating bad tautologies in the subsentential structure of (T^3). So observing this rule means that no such inferences over all of the variables together contained in the conjunction of (T^3)'s antecedent is permissible. Then (T^3)'s subsentential application of logical connectives merely displays the structure of transitive preference over pairwise comparisons, without permitting any further logical inferences over their individual variables across multiple occurrence of *P*.

However, we do not need to be able to perform any such inferences across the individual variables contained in (T^3) independent of the sentential *P*-functions in which they are contained. Nor do we need to verify (T^3) as a truth of logic. All we need (T^3) to be is consistent, and all we need to be able to do is give its variables *x*, *y* and *z* an ordinal ranking on a utility scale. But a conscious and intentional chooser's memory of the relation of *x* to *y* while she is ranking *y* over *z* is what enables her simultaneously to form and apply the concept of *x*'s ranking superiority both to *y* and to *z*; of *z*'s ranking inferiority both to *y* and to *x*; and thereby to infer from her selection behavior at t_1 and t_2 that she prefers *x* to *z*. Together with (Con), above, it therefore enables her to weakly order *x*, *y* and *z* relative to one another on a utility scale. So I think the right response to the dire consequences described above of violating the No-Trespass Rule in (T^3) is to just not violate the rule. We will see shortly that this advice has no untoward implications for our answers to questions (i) or (ii) of Section 3.

8. The Variable Term Calculus: Subsentential Predication

So far I have suggested some notational revisions to Savage's system designed to enable us to represent the language of preference within the familiar constraints of the Boolean connectives. Essentially these have amounted to embedding and expanding within the place conventionally held by variable terms some familiar operations of sentential logic on the variables to which an *n*-adic predicate ordinarily is ascribed; this is why I describe these

proposed revisions as constituting a variable term calculus. In order for the variable term calculus to represent an ordinal ranking in an intuitively acceptable way within the constraints imposed by the Boolean connectives, certain further notational revisions, familiar from predicate logic, need to be introduced.

Let *A* be a two-place predicate that denotes the "above" relation. Then

(1) *Pw(Axy)*

states that w ranks (or prefers) *x* above *y*. Notice first that (1) avoids begging the questions raised in Section 2 against Savage's assumptions about the numerical commensurability of *x*, *y* and *z*. I may rank veggies above rice without being committed to any sense in which veggies are more than beans (other than the unhelpful sense in which they perhaps *mean more* to me). A noncommittal stance toward numerical commensurability is a virtue in an ordinal ranking of alternatives.

(1) introduces the possibility of conceiving not only variable terms but predicate letters – and, if needed, quantifiers as well – as subsentential constituents that can be nested within other predicates that govern entire sentences, such that the scope of the outermost is the entire sentence whereas the scope of one enclosed within the brackets is the variable term(s) enclosed within sub-brackets to the right. Call the outermost *governing predicates*. In this discussion, *P* would be a governing predicate. *A*, like any predicate letter whose scope is a variable term or relation among some but not all variable terms in the sentence, would exemplify what I shall call a *subsentential predicate*.

The same constraints on linguistic interpretation mentioned in Section 4, above, apply here. So, for example,

(2) *(Axy)*

is not a sentential proposition but rather a constituent of one that says merely "…*x* above *y*…." And similarly, the interpretation of (2) will depend on the context and intentional operator that modify it. (1) demonstrates the interpretation of (2) in a sentence asserting a preference ranking. In a sentence asserting a desire, (2) will be spelled out as the desire *for x above y*, or *for x to be above y*, or *that x be above y*. Or a sentence may incorporate (2) similarly as the object of intending *x above y*, or intending *x to be above y*, or *that x be above y*. Or (2) may express an agent's perception of *x as being above y*, or her believing *x to be above y*; and so on. Within this stipulation, more fine-grained semantic ambiguities are resolvable with the provision of additional context.

Correspondingly,

(3) *(∃y)(x)Axy*

would merely mention "… a(n existing) *y* such that all *x*s above it…" rather than asserting sententially that there is such a one. Thus I might believe all *x*s to be above a *y*, or intend any *x* to be above an existing *y*. Or I might rank any *x* above an existing *y*, and express it thus:

(4) *Pw[(∃y)(x)Axy]*

It is to be hoped that the general idea is clear: it is to do for subsentential constituents with predicate and quantificational logic what I earlier suggested we do with sentential logic, with the same rules and restrictions, plus those peculiar to quantificational inference.

One benefit of this approach is that it permits a restatement in quantificational terms of Savage's original conception of ordinality (O) (Section 2, above) that captures what we need from the original:

(O') *{[Pw(x.~y).Pw(y.~z).→Pw(x.~z)]. [Pw(x.~y).→~Pw(y.~x)]}*
→ (∃z) [Pw(Axy.Ayz)]
(Ordinality)

(O') says that if *w*'s preferences among *x*, *y*, and *z* are transitive and asymmetric, then *w* ranks *x* above *y* and *y* above an existing *z*; i.e. that *w*'s ordering of *x*, *y*, and *z* has a lowest-ranked member and so constitutes a well-ordered triad. (O') enables us to answer the objections raised to Savage's conception of a simple ordering raised in Section 2 *by avoiding any suggestion as to the selection criteria on which pairwise comparisons are based*. As predicted, this notation sacrifices the streamlined elegance of Savage's measurable and uniform rendering. But as promised, it also avoids begging the question as to what those selection criteria are.

A second benefit of introducing subsentential predication into the variable term calculus is that it allows us to symbolize a noncyclical solution to Gertrude's choice problem as described in Section 2. Recall that Gertrude preferred chocolate ice cream to vanilla for its sweetness, vanilla to coffee for its taste, and coffee to chocolate for its texture; that she continued to prefer each flavor of ice cream for one of its properties, and also something that was not that flavor for a different property; hence that her choice dilemma could not be described as a typical cyclical ranking. Recall also that the appearance of cyclicity in Gertrude's preference ranking arose out of her failure to rank independently the relevant properties themselves – sweetness, taste, and texture – of the alternatives she confronted. In the notation of subsentential predication I am proposing, the inconsistency arising out of Gertrude's continuing preference for each flavor of ice cream for one of its properties, and also something that is not that flavor for a different property can be more accurately expressed. Let individual variables *a*, *b*, and *c* denote chocolate, vanilla, and coffee ice creams respectively. Then Gertrude's preference for each of chocolate, vanilla, and coffee for one of its properties, and also something that was not that flavor for a different property is symbolized as follows:

(5) *Ps(a.~a) . Ps(b.~b) . Ps(c.~c).*

That is, Gertrude's preference as originally stated violates (Irr) and hence is formally self-contradictory. And the right way of ironing out this self-contradiction is for Gertrude to rank independently the relevant properties themselves – sweetness, taste, and texture – of the alternatives she confronts. Let the predicate letters *S*, *T* and *R* denote sweetness, taste and texture respectively. Then Gertrude's task is to consider whether perhaps

(6) *(∃P)(S)(T)(R)[Ps(S.~T) . Ps(T.~R) . Ps(S.~R)]*

for the three alternative flavors she is offered. If there is, indeed, a preference P such that Gertrude prefers sweetness to taste and taste to texture in ice cream, then with the aid of (O'), above, *S*, *T*, and *R* can be ordered thus:

(7)*{[Ps(S.~T).Ps(T.~R).→Ps(S.~R)]. [Ps(S.~T).→~Ps(T.~S)]}→ (∃R) [Ps(AST.ATR)]*

Hence Gertrude's ordering of sweetness, taste, and texture has a lowest-ranked member – texture – and so constitutes a well-ordered triad. With this ordering of properties, Gertrude can now produce a transitive ordering of the three flavors of ice cream with which she is confronted that respects the variety of properties that determines that ordering:

(8) *Ps(Sa.~Tb).Ps(Tb.~Rc).Ps(Sa.~Rc)*.

Another, potential application of such a property ordering would treat numerically nondeterminate degrees of probability or Bayesian confidence as properties of alternatives the ranking of which might similarly modify the ranking of those alternatives.[21]

9. De Jongh and Liu's Constraint-Based Analysis of Strict Preference

With the above property ordering and the account of subsentential predication in which it is embedded in hand, I now examine briefly a competing analysis of strict preference that begins with the same intuitions as mine about first-order logical formulations of it, but introduces predicate letters in advance of the intensional apparatus I have proposed so far. De Jongh and Liu approach the formulation of the preference relation through the lens of optimality theory in linguistics.[22] Sometimes a uniquely optimal solution – a single and singularly correct speech act appropriate to the circumstances – cannot be produced by the grammatical theory in question. In this case, optimality theory first engenders a set of alternative solutions: for example, the set A consisting in

> {"Glad to meet you.", "Hey, man!", "Yes, well, hmmm …", "It's good to meet you.", "How nice …?", "Howdy!", "It is a privilege to make your acquaintance.", "Yo!", "How do you do?", "Charmed, I'm sure."}.

A set of conditions or constraints, strictly and lexically ordered according to their importance, is then applied to these alternatives, and the alternative that best satisfies conditions imposed earlier in the sequence is stipulated to be a uniquely optimal solution. Thus the ordering of alternatives is fixed by their more or less successful satisfaction of the constraints. For example, the set C consisting in

> {expresses respect for the eminent personage to whom one is being introduced, is acceptable at a formal foreign embassy dinner, is uttered at a first meeting among strangers, puts both speakers at ease, establishes relations of casual familiarity}

picks out "It is a privilege to make your acquaintance" as a uniquely optimal solution relative to C.

Roughly speaking, then, a *constraint* is a linguistic formula that functions logically in much the same way as does a predicate in quantificational logic, i.e. it is ascribed to a variable and satisfies the law of non-contradiction for predicates *(x)~(Fx . ~Fx)* – or, as De Jongh and Liu

express it, "either the constraint clearly is true of the alternative or it is not." Hence for purposes of this exposition, we can think of their concept of a constraint as a certain kind of predicate.

De Jongh and Liu are interested in the way in which the imposition of constraint predicates engenders a preference ordering among all the alternatives. Hence their approach treats constraint predicates as giving rise to a preference ordering among alternatives, but also presupposes a strict ordering among those predicates themselves. Later in the discussion, they then go on to examine the way in which introduction of the belief operator in doxastic logic offers new ways of thinking about preference change; it must be emphasized that this is their primary concern. But my interest here is confined to De Jongh and Liu's conceptualization of the relation between preference alternatives and the predicates that are argued to order them.

De Jongh and Liu define a constraint sequence as a finite, strictly ordered sequence of constraints C_1, C_2, … C_n, each of which is predicated of exactly one free variable x, such that

(1) $C_1x > C_2x \ldots > C_nx$

and, for example, $C_1 . \sim C_2 \ldots . \sim C_m$ is preferable to $\sim C_1 . C_2 \ldots . C_m$; and $C_1 . C_2 . C_3 . \sim C_4 . \sim C_5$ is preferable to $C_1 . C_2 . \sim C_3 . C_4 . C_5$. They then define a strict preference for x over y ***Pref (x,y)***, given a constraint sequence C with n members, as follows:

(2) $Pref_1(x,y) =$df. $C_1x . \sim C_1y$

(3) $Pref_{k+1}(x,y) =$df. $Pref_k(x,y) \vee [(C_1x \equiv C_1y) . \ldots . (C_kx \equiv C_ky)^{23} . C_{k+1}x . \sim C_{k+1}y]$, $k<n$

(4) $Pref(x,y) =$df. $Pref_n(x,y)$.

Similar in logical structure to my $Pw(x.\sim y)$, (2) intuitively defines preference for x over y with respect to the first C in n as the case in which that first and lexically prior constraint predicate C_1 holds true of x and not of y. On that basis, (3) then defines preference for x over y with respect to subsequent constraints C_{k+1} in n as the case in which either x is preferred to y with respect to any arbitrarily selected constraint C_k in n; or else x and y are equivalent for any C_k in n and x is preferred to y with respect to any subsequent constraint C_{k+1}. In (3), either earlier constraints select a single preferred alternative; or else the two alternatives are constraint-equivalent and subsequent constraints select the same single preferred alternative. (4) defines strict preference for x over y as the case in which these stipulations hold for the last constraint in n. To illustrate how to use this definition to move from constraints to preference, De Jongh and Liu offer an example in which Alice has the constraint sequence $Cx > Qx > Nx$, such that Cx means "x has a low cost," Qx means "x is of good quality," and Nx means "x is in a nice neighborhood;" and must choose between two houses d_1 and d_2 with the properties Pd_1, Pd_2, $\sim Qd_1$, $\sim Qd_2$, Nd_1, and $\sim Nd2$. Since d_1 and d_2 both bear P and lack Q, d_1 and d_2 are ordered on the basis of the last constraint Nx in the sequence, which determines Alice's strict preference of d_1 over d_2, i.e. $Pref(d_1,d_2)$.

De Jongh and Liu's definition is very useful for the case in which constraint predicates are ascribed to states of affairs that include properties additional to those for which one has an identified strict preference ranking, and also to those for which some ranked property fails to hold. However, why these should enter into a definition of strict preference is unclear. If both d_1

and d_2 have P and lack Q, then neither P nor Q enter into Alice's strict preference ranking. P does not because Alice gets P in either case (perhaps Px is "has a roof"); and Q does not because she fails to get Q in either case. Then the properties that she is actually required to strictly order are Cx and Nx; this can be done with a pairwise comparison plus the usual conditions (asymmetry, irreflexivity, transitivity), in the way suggested above (Section 8.(6) and (7)). De Jongh and Liu are interested in other varieties of order besides strict ones, and correspondingly non-strict conceptions of preference. Their definition of strict preference is meant to extend to these other varieties, as well as to belief contexts; but is less intuitively plausible for the standard case on which their analysis is based.

Moreover, in De Jongh and Liu's notation, the heavy lifting in ordering preference alternatives is driven by the predicates that modify them, rather than – as in mine – the first-order logical structure of strict preference itself, whether this orders preference alternatives or the predicates ascribed to them. But in order to do this heavy lifting, De Jongh and Liu's constraint predicates must be given a strict and linear ordering antecedently, which reintroduces the connective problem that the concept of a constraint had seemed to dissolve. De Jongh and Liu use the mathematical connective ">" for this purpose, as is conventional. Through De Jongh and Liu's definition of strict preference, the linear ordering of constraint predicates given by ">" then determines the ordering of preference alternatives to which those predicates are ascribed. But this relation, and therefore the choice procedure in which it is nested, is subject to all of the objections I have already raised in this chapter to the putative extensionality of standard decision-theoretic notation. Using the foregoing account of subsentential predication, by contrast, the work of ordering choice alternatives can be performed just as well by the same quantificational apparatus and sequence of Boolean connectives as is used to order the predicates ascribed to them, while at the same time avoiding these objections. Alice's preference for house d_1 over d_2 can be written as follows, where Rx is "has a roof," Qx is "is of good quality," and Nx is "is in a good neighborhood":

(5) $Ps(Rd_1 \vee Rd_2).\ Ps(\sim Qd_1 \vee \sim Qd_2).\ Ps(Nd_1.\sim Nd_2)$

On my account, it is not necessary to stipulate the problematic, antecedent mathematical ordering of constraints as a precondition for applying the subsequent first-order logical definition of preference, as De Jongh and Liu do, because the standard Boolean connectives and quantificational laws of first-order, classical predicate logic are all we need.

10. The Intensionality of Genuine Preference

Conjointly, (Asy), (Con), (Irr), (T^3), and (O') and constitute a conceptual truth about what it means for someone to have a genuine preference. Its status as a preference is stable relative to the rejected alternative (Asy); it has been compared to all other alternatives in the given set (Con); it satisfies horizontal consistency, i.e. is not self-contradictory (Irr); it is consistently preferred to all other alternatives in the set (T^3); and it is well-ordered relative to the least member of the set

(O'). Together these five criteria insure that something is a genuine preference if it is consistent both with itself and with each of the other alternatives to which it is preferred.

Notice that the suggested notational revisions do not require any sacrifice of content in the expression of probabilistic axioms. For example, the Von Neumann-Morgenstern independence axiom discussed in Volume I, Chapter IV.1.2,

(Ind) if F > G and $0 < p < 1$ then $F(p) + H(1 - p) > G(p) + H(1 - p)$ for any H in the set S of all probability distributions or gambles on a set of outcomes

can be rewritten in the variable term calculus as follows, where *x*, *y*, and *z* are alternatives and "*Sz*" denotes any *z* in the set S of all probability distributions etc.:

(Ind') $(z)\{(Sz \rightarrow \{[Pw(x.\sim y).(0 < p < 1)] \rightarrow \{Pw[x(p) + z(1 - p)].\sim[y(p) + z(1 - p)]\}\}\}$.

One advantage of this notation is that it confines the mathematical connective "<" to the extensional entities while substituting the standard sentential and quantificational connectives for the intensional ones.

Now to situate these considerations relative to the rationality conditions proposed in the preceding chapter. That chapter's requirement of horizontal consistency required that for any agent's set *S* of concepts of things and properties $c_1, c_2, c_3, \ldots c_n$, and rationally intelligible things or properties $t_1, t_2, \ldots t_n$ assigned to individual variables $a_1, \ldots a_n, b_1, \ldots b_n$,

(HC) $(\sim\exists x)(x.\sim x)$.

i.e. we must conceive any such c_i as self-identical, or nonself-contradictory. Where c_i is the concept of a genuinely preferred alternative t_i, (HC) secures the mutual logical consistency of all preference alternatives and their properties simultaneously intelligible to me at a particular moment with all the other things and properties equally intelligible to me at that moment. Satisfaction of the additional five criteria just discussed ensures that my choice will constitute a genuine preference that is also horizontally consistent with the other beliefs and preferences constitutive of my perspective over some arbitrarily selected stretch of time (i.e. subject to the caveats about changes in agent preferences determined by personal growth and character development over time emphasized in Chapter II.4).

Similarly, the requirement of vertical consistency discussed in the preceding chapter secures the intensionality of a genuine preference by anchoring it in my perspective as an experience I have. Where thing or property t_i is a genuinely preferred alternative, then given an individual variable *a* to which t_i is assigned, and terms *F* and *G* with the extensions *P* and P^1 respectively,

(VC) $Fa \rightarrow [(x)(Fx \rightarrow Gx) \rightarrow Ga]$.

Applied to preference alternatives themselves, (VC) states that if preferred alternative t_i is a *P*, then if all *Ps* are P^1s, then t_i is a P^1 as well. In particular, whatever other properties t_1 has, all such preferred alternatives are objects of my experience; therefore t_1 is, too.

However, an agent's preferring one alternative to another is itself a triadic relational property that holds among the agent and the two alternatives in question. Hence (VC) has deeper implications for genuine preferences. Vertical consistency for genuine preferences implies

that if an agent *s* prefers *a* to *b*, then where one can prefer an alternative to another only in light of some further triadic relational property that holds among oneself, the preferred alternative and the rejected alternative, then *s* bears that property in relation to those two alternatives as well: If *s* prefers *a* over *b*, then if preferring one alternative to another implies *Q*-ing one alternative to another, then *s* also *Q*s *a* over *b*. More precisely,

$$(VC^P)\ Ps(a.\sim b) \rightarrow \{(w)(\exists x)(\exists y)[Pw(x.\sim y). \rightarrow Qw(x.\sim y)] \rightarrow Qs(a.\sim b)\}.$$

Some important candidates for *Q* include intending to bring about, remembering, and deliberately furthering. This means that all genuine preferences as such, regardless of their objects or the properties those objects may have, necessarily bear certain further properties in common: something that is a preferred alternative is also an intentional object, an object of consciousness, and an object of deliberate action. An object of desire, on the representational analysis offered in Volume I, Chapter II can be a genuine preference only to the extent that it satisfies (VC^P).

Perhaps most importantly, we can now see how (T^3) itself instantiates (VC):

$$(VC^T)\ Ps(a.\sim b).Ps(b.\sim c) \rightarrow$$
$$\{(w)(x)(y)(z)[[Pw(x.\sim y).Pw(y.\sim z)] \rightarrow Pw(x.\sim z)] \rightarrow Ps(a.\sim c)\};$$

and therefore that transitive preferences are not only genuine preferences, but indeed genuine preferences that fit with vertical consistency into an agent's perspective. (VC), (VC^P) and (VC^T) most inclusively require, then, that genuine preferences also bear the self-consciousness property, i.e. that all such preferences be objects of experiences the chooser has. Hence genuine preferences as defined in the proposed variable term calculus are integrated into an agent's perspective as some among many other experiences that also include thoughts, beliefs, perceptions and emotions. As we have already seen in the preceding chapter, this requirement, together with that of horizontal consistency, secures the rational intelligibility and logical consistency of a chooser's preference; and the self-determining agency of that chooser. I argue in Chapter VIII.7 below that logically consistent preferences – those that satisfy the requirements of horizontal and vertical consistency – thereby terminate the infinite regress problem we saw in Volume I, Chapter VIII.2 was of such concern for Humeans. Thus the significance of vertical consistency for the concept of a genuine preference is that it makes this the concept of *some one psychologically consistent subject's* genuine preference. Henceforth when I speak of something's satisfying the criteria of horizontal and/or vertical consistency *over time*, I shall mean to denote its satisfaction of (Asy), (Con), (Irr), (T^3), and (O') in addition to its satisfaction of (HC) and (VC). However, as we have seen in Chapter II.4, not all preferences necessarily satisfy all or even most of these conditions; which ones do depend on empirical considerations.

For Kant, all classical logic was intensional because it structured the most fundamental categories of our cognition and experience. Classical logic for Kant mirrors the limits of logical possibility in empirical reality for the same reason Euclidean geometry mirrors the limits of spatiotemporal experience in empirical reality: both are necessary and constitutive mental preconditions for experiencing an empirical reality at all, and both presuppose the agent's

perspective the structure and outer limits of which they circumscribe. Because predicate logic provides structure and constraints to the objects of possible experience within an agent's perspective, it first-personally expresses the logical limits of that experience in general. By contrast, the special case of it developed here as a variable term calculus first-personally expresses the logical limits of consistent preference in particular.

Now I contended in Section 1 that formulation of a rule of transitivity for preferences requires that the intensional conditions under which it holds be fully spelled out. I also argued that orthodox decision-theoretic formalizations were at a loss to do this because (1.9) holds neither for an actual empirical chooser, nor an ideally rational chooser under conditions of uncertainty, nor for an ideally rational chooser under conditions of full information. With the aid of the variable term calculus we are in a better position to spell out the intensional conditions under which the rule of transitivity holds for preferences. (Asy), (Con), (Irr) [or (HC)], (T^3), (O'), and (VC) conjointly formalize the two necessary conditions of conscious and intentional choice listed in Section 2:

> (a) A chooser must be able to form and apply consistently through time the concept of a thing's ranking superiority – and therefore some other thing's ranking inferiority – over a series of pairwise comparisons; and
>
> (b) she must remember the relation of the two alternatives she is presently ranking to the third she is not.

(a) is satisfied just in case (VC), (Irr) [or (HC)], (T^3), and (O') are; and (b) is satisfied just in case (VC), (Asy) and (Con) are. That is, a chooser forms and applies consistently through time the concept of an alternative's ranking superiority over a series of pairwise comparisons if and only if that alternative is subsumed under that concept, is not self-contradictory, is consistently preferred to all other alternatives in the set, and is well-ordered relative to the least member of the set. A chooser remembers the relation of the two alternatives she is presently ranking to the third she is not if and only if the status of the preferred alternative as preferred is identifiable (i.e. by the concept of ranking superiority), stable relative to the rejected alternatives, and has been compared to the other alternatives in the set. Satisfaction of these criteria neither requires nor precludes a chooser's empirical actuality, ideality under uncertainty, or ideality under conditions of full information. Nor does it require that a chooser's preferences be epistemically transparent. All it requires is that the chooser have a genuine – that is, a rationally considered – preference, and not merely a sudden and mercurial yen.

11. The Consistency of Savage's Simple Ordering (T^3)

Now, with the aid of *P* and the six criteria of genuine preference that define it, we can state in detail some ways in which a cyclical ranking not only fails to express a genuine preference, but in fact expresses a logical inconsistency. Let us now use the proposed notation to restate the questions about Savage's concept of a simple ordering posed in Section 3:

> (3.i') Can $Pw(x.\sim z)$ and $Pw(z.\sim x)$ both be true together?

The answer is clear at a glance: not without violating (Asy).

Then define a cyclical ordering (C') of alternatives *x*, *y*, and *z* as follows:

(C') *Pw(x.~y).Pw(y.~z).Pw(z.~x)*

If (T^3) were a tautology, good or bad, (C') would be a logical impossibility. This would be a bad thing, since actual agents do sometimes produce cyclical orderings. Observing the No-Trespass Rule circumvents this evil, while preserving (C')'s susceptibility to logical requirements both sententially and subsententially. In light of it, we can now rephrase our second question from Section 3:

(3.ii') Can (T^3) and (C') both be true together?

The following derivation suggests that they cannot:

(T^3) *Pw(x.~y).Pw(y.~z).→Pw(x.~z)*	Premise
(C') *Pw(x.~y).Pw(y.~z).Pw(z.~x)*	Premise
(1) *Pw(x.~y).Pw(y.~z)*	(C')
(2) *Pw(x.~z)*	(T^3), (1)
(3) ∴ *Pw(x.~z).Pw(z.~x)*	(2), (C')

Step (3) shows that (T^3) in conjunction with (C') violates (Asy). If we then bring in (Asy) as an additional premise,

(A) *Pw(x.~z).→~Pw(z.~x)*	Premise
(4) *~Pw(z.~x)*	(Asy), (2)
(5) *Pw(z.~x)*	(3)
(6) ∴ *Pw(z.~x).~Pw(z.~x)*	(4), (5)

we see that (T^3) and (A) together, when joined with (C'), generate a straightforward logical contradiction. This explains why (T^3) and (C') cannot logically be true together.

Finally, we can show how (C') violates the requirements of logical consistency that (T^3) satisfies, by recurring to the case considered in Volume I, Chapter IV.3.1, of Rex, who has and applies the concept of ranking superiority to *z* – hence gives a cyclical ordering – because he has forgotten the relation of *x* and *y* to *z* established by his two previous rankings. I argued that in that case, every alternative is preferred to every other, hence that none of the three is superior in ranking to any of the others, and so none superior to *x*. I concluded that Rex's application of the concept of some one thing's ranking superiority to *z* at t_3 therefore had involved him in a logical contradiction, i.e. that *z* both was and was not preferred to *y*. However, the restrictions of conventional preference notation gave us no way to express this conclusion formally. With the aid of the variable term calculus I have suggested here, we are now in a better position to express formally the thought that a cyclical ranking is logically contradictory. Rex ranks *x*, *y* and *z* as follows:

(7) t_1: *Pw(x.~y)*

(8) t_2: *Pw(y.~z)*

(9) t_3: *Pw(z.~x)*

From (8) and (9), (T^3) permits the inference to

(10) *Pw(y.~x).*

From (9) and (7), (T^3) permits the inference to

(11) *Pw(z.~y).*

And from (7) and (8), (T^3) permits the inference to

(12) *Pw(x.~z).*

This much simply translates Savage's notation into mine. But before, in Volume I, Chapter IV, we could state the crucial conclusion to logical inconsistency only in natural language and not symbolically in Savage's notation. We can now, however, state it symbolically in the variable term calculus. From (7) and (10), (T^3) permits us to infer

(13) *Pw(x.~x),*

which violates (Irr) [or (HC)]. We can now see more clearly that a cyclical ordering is logically self-contradictory.

Recall why it was useful to establish this. Volume I, Chapter IV argued that the prevailing interpretation of the utility maximization model of rationality as having universal application implied that it was either vacuous or logically inconsistent. The example of Rex (inter alia) was invoked to demonstrate that this implication could not be deflected by imposing purely decision-theoretic consistency constraints such as (T) on a preference ranking, because any apparent violation of (T) could be interpreted as absentmindedness, or a mere change of mind on the part of the chooser. A chooser such as Rex might then be accused of psychological inconsistency. But there was nothing inherent in the unreconstructed utility maximization model of rationality that requires a rational agent to be psychologically consistent, and no resources within the conventional constraints of this model for inferring from mere *psychological* inconsistency any violation of (T) – or, therefore, inconsistency in any more robust sense.

I concluded that in order to show a cyclical ordering to be an inconsistent one – and therefore the model itself to be more than a mere tautology, this model needed to be subsumed under the rubric of a broader conception of rationality – the traditional one based in sentential and quantificational logic that Kant embraced – that possessed the formal resources to analyze it accordingly. In this chapter I have proposed a variable term calculus as a way of doing this. This approach subordinates the utility maximization model of rationality to the more general and universal requirements of classical logical consistency, and so divests it of its pretensions to universality of application. But in exchange, it receives the status of a genuine, disconfirmable theory. Under the umbrella of a Kantian model of rationality, utility theory becomes more than a meaningless truism about always doing what we most want to do.

Endnotes to Chapter III

[1] I use the term "classical logic" to apply inclusively to the logics of sentential propositions that are categorical in form and declarative (i.e. indicative) in mood. This covers both the sentential calculus, in which sentential propositions s_1, s_2, s_3, … are symbolized by sentence letters P, Q, R, …; and also the quantificational calculus, in which categorical declaratives are symbolized in combinations of variable terms x, y, z, … and predicate letters F, G, H, … having the extensions P, P^1, P^2, ….

[2] That orthodox decision theory does, in fact, treat ">," "≥," etc. as extensional connectives is not seriously open to doubt. See, for example, John Von Neumann and Oskar Morgenstern, *Theory of Games and Economic Behavior* (Princeton: Princeton University Press, 1990), 18-19, fn. 3 for a treatment of ">" and "<" as having the same notational status as "=". For the straightforward appropriation of such connectives from the mathematical into the decision-theoretic context for purposes of formally defining a simple ordering, see Leonard Savage, *The Foundations of Statistics* (New York: Dover Publications, Inc., 1971), 18-19; for purposes of formally defining utility-maximization, see R. Duncan Luce and Howard Raiffa, *Games and Decisions* (New York: John Wiley and Sons, Inc., 1957), 15; and for purposes of formally defining majority decision, see Amartya K. Sen, *Collective Choice and Social Welfare* (San Francisco: Holden-Day, Inc., 1970), 71.

[3] Robert Paul Wolff proposed this argument to me (personal communication, August 17, 2001).

[4] Savage, *The Foundations of Statistics*, *ibid*., 21.

[5] *Ibid*., 17-19.

[6] However, it should be noted that not all preferences are comparative. A gentleman who prefers blondes, for example, is one who is always physically attracted to blondes – not one who always selects blondes whenever offered a pairwise comparison between a blonde and everybody else.

[7]This proposal was originally suggested to me by Henry Richardson some years ago (personal communication, October 2, 1987).

[8] Not only are there some noncomparative preferences, such as that of gentlemen for blondes. There are even some intensional objects of preference that cannot be *interpreted as the outcome* of a pairwise comparison. Here is an example of one that cannot: Myrtle prefers blonds. But given a series of pairwise comparisons between any particular blond and any non-blond potential suitor, Myrtle always selects the one who has the best sense of humor.

[9] Fenrong Liu and Dick de Jongh arrived at a similar interpretation of strict preference independently. They proposed it in their "Optimality, Belief and Preference," delivered to the *Models of Preference Change* Workshop, Freie Universität Berlin, 15 September 2006; and published in *Proceedings of the Workshop on Rationality and Knowledge*, Ed. Sergei Artemov and Rohit Parikh, Eds (ESSLLI, 2006). I discuss their proposal in Section 9, below.

[10] *Op. cit.* Note 2, Luce and Raiffa, *Games and Decisions*, 302.

[11] See his *Rationality and Dynamic Choice* (New York: Cambridge University Press, 1990), Chapter 3 and Footnote 2, pp. 269-270. I am indebted to the insights in this book at very many points in this discussion, even – and perhaps most particularly – where my own view diverges from them.

[12] Mark Kaplan, *Decision Theory as Philosophy* (New York: Cambridge, 1996), 23 – 32; or DTP. References to this work are henceforth paginated in the text.

[13] Mark Kaplan, personal communication of July 10, 2003.

[14] *Ibid*.

[15] (6.i – iii) are the Kolmogorov axioms of probability. Here and throughout I have condensed Kaplan's lucid exposition for the sake of brevity.

[16] To his credit, Kaplan nevertheless explores this alternative in Kaplan (*op. cit.* Note 12), pages 12-13.

[17] Quine describes this method in *Methods of Logic*, 3rd Edition (New York: Holt, Reinhart and Winston, Inc., 1972), Chapter 6, "Consistency and Validity."

[18] In the following discussion I rely on Richard Jeffrey's exposition in *The Logic of Decision*, 2nd Edition (Chicago: University of Chicago Press, 1983), Chapter 9: "Existence: Bolker's Axioms." I have also learned from studying Ethan Bolker's terser and more demanding treatment in "A Simultaneous Axiomatization of Utility and Subjective Probability," *Philosophy of Science 34*, 4 (December 1967), 333-340; and "An Existence Theorem for the Logic of Decision," *Philosophy of Science 67* (Proceedings 2000), S14-S17.

[19] *Op. cit.* Note 2, 17.

[20] *Ibid*., 21.

[21] For a thought-provoking discussion of this possibility, see Mark Kaplan, "Decision Theory and Epistemology," Section III, in Paul K. Moser, Ed., *The Oxford Handbook of Epistemology* (New York: Oxford University Press, 2002).

[22] Op. cit. Note 9.

[23] De Jongh and Liu substitute $Eq_k(x,y)$ for $(C_1x \equiv C_1y) . \ldots . (C_kx \equiv C_ky)$ for brevity. I restore the original sentence in order to expose the structure of their definition. I also translate their notation for the Boolean connectives into mine for purposes of comparison.

Chapter IV. McClennen on Resolute Choice

Chapter III developed the concept of a genuine preference to anchor proposed modifications in canonical decision theory. These demonstrated how the utility-maximizing model of rationality with which canonical decision theory is traditionally identified is in fact only a special case of a more comprehensive model of rationality to which canonical decision theory, suitably modified, is entirely adequate. In this chapter, by way of applying this conclusion, I examine Edward McClennen's concept of resolute choice (thus redeeming the promissory note I issued in Volume I, Chapter III.1), which he develops within the constraints of the unreconstructed utility-maximization model. I show that McClennen's notion of resolute choice is justified independently of utility-maximization, and offers an incentive to action that is independent of utility-maximization, because it is in fact materially equivalent to my Kantian concept of a genuine preference. Decoupled from the issue of whether or not utility is maximized, McClennen's model in effect imposes a nomological requirement on rational choice that identifies the concept of resoluteness as what Kant would call a law. In this regard, McClennen's pragmatic model of resolute choice succeeds, despite McClennen's own resistance, where Kant's rationalist model fails: in deriving the obligation of promise-keeping from the concept of reason. It thereby suggests a new, intrapersonal solution to the free rider problem. Finally, the concept of resolute choice implies an account of moral emotion that is independent of interpersonal dynamics or social conditioning.

Section 1 embeds McClennen's concept of resolute choice within his project of providing a utility-maximizing justification for both a commitment and a psychological disposition to behavior guided by rules. Section 2 discusses McClennen's contrast between such a commitment and the myopic choice that a strict interpretation of (U) seems to require, in which cost-benefit deliberation about a particular choice is isolated both from consideration of choices made in the past and from projections about choices to be made in the future. Section 3 describes two strategies for circumventing the disadvantages of myopic choice: precommitment and sophisticated choice, and explicates and supplements McClennen's reasons for rejecting both as irrational. Section 4 introduces McClennen's analysis of resolute choice, and shows how it both resolves problems of personal continuity that myopic choice engenders, and also integrates the chooser psychologically by coordinating successive temporal stages through the commitment to rule-guided behavior. Section 5 shows that the viability and effectiveness of resolute choice as McClennen describes it does not depend on the utility-maximizing considerations by which he justifies it, and Section 6 demonstrates its material equivalence to my concept of a genuine preference. Section 7 compares McClennen's concept of resolute choice with Kant's concept of law-governed self-determination, and argues that the commitment to promise-keeping can be successfully derived from the former if not from the latter. Finally Section 8 shows how, when disjoined from considerations of utility-maximization, the concept of resolute choice offers a solution to the free rider problem that goes well beyond traditional conceptions of it as a problem

of interpersonal coordination; and thereby offers an intrapersonal foundation for normative moral theory.

1. McClennen's Project

McClennen aims to provide what he describes as a "consequentialist" defense of a certain kind of rule-guided behavior under certain specific circumstances. Following Peter Hammond,[1] he defines *consequentialism* as the view that "choice of an action is acceptable if and only if the consequences of that action are maximally preferred by the agent – if and only if the agent chooses so as to maximize with respect to his preference ordering over consequences [RDC 83]." [2] This definition of consequentialism is recognizable as a variant on the minimalist formulation of (U) in Volume I, Chapter III.1, that if a rational agent acts, she maximizes utility. Both locate utility-maximization 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. So as to mark the distinction between this concept and the moral concept of consequentialism examined in Volume I, Chapter V, I hereafter substitute the term *utility-maximizing* for "consequentialist." Thus I describe McClennen's project as a utility-maximizing defense of rule-guided behavior; and sometimes refer to (U) where McClennen or Hammond would refer to consequentialism.

McClennen's aim is also distinct from a defense of rule-utilitarianism, in that it does not make the strong assumption that agents are motivated to follow certain rules by a benevolent desire to maximize total or average utility using an overall social welfare function. Instead, McClennen's project is the weaker and more inclusive one of defending "rule consequentialism, in which the notion of a rational commitment to extant rules has a central place [PRR 258]," and the relevant notion of rationality is the utility-maximizing one just defined, regarding "what is advantageous to a given person from each of the relevant temporal points in a series of choices to be made over time, or mutually advantageous to a set of persons who find themselves faced with a problem of interdependent choice [PRR 216, fn. 13]." Although McClennen considers both the intrapersonal and the interpersonal cases, I confine my attention to the former.

By *rule-guided* behavior, McClennen means action not only in conformity with certain rules, but action that in addition includes an intentional pro-attitude toward the rules themselves, such that an agent's reason for conforming his behavior to a particular rule is that the rule requires it. Thus rule-guided behavior includes a commitment – or, as McClennen defines this – a psychological disposition to follow the rule because the rule requires it [PRR 211]. This conception of rule-guidedness is consonant with Kant's thesis that "[o]nly a rational being has the capacity to act *in accordance with his representation* of laws – that is, in accordance with principles, i.e. a *will* [Ak. 412; italics in text]." We shall see that the mutual interdependency Kant claims between rule-guidedness and will is an interdependency that is central for McClennen as well. McClennen means to show that a psychological disposition to rule-guided behavior could arise, not only from involuntary socialization or hard-wired biological drives, but also from rational

deliberation aimed at maximizing utility. The basic idea is that under certain circumstances an agent might be disposed to guide his behavior according certain sorts of rules in order to maximize freedom, flexibility, or scarce resources because he understands that violating the rule would be costly of these things. McClennen does not claim to offer a defense of pervasively or consistently rule-guided behavior on utility-maximizing grounds. This is as it should be, since as we acknowledged in Volume I, Chapter IV.5, guiding one's behavior according to rules will not always maximize utility in the unreconstructed sense of (U). Hence the psychological disposition to so guide one's behavior would be activated only by the agent's deliberative conclusion that doing so would maximize utility.

2. Myopic Choice

McClennen's project is a challenging one because it seems to conflict with the very idea of utility-maximization as defined above. Because it focuses on the choice of a particular action in light of its consequences, (U) would seem to restrict considerations influencing choice among pairwise-compared alternatives to what will maximize utility from the moment of that choice forward in time, without regard either to choices made at earlier points in time that might be expected to have some impact on this one, or to choices to be made later in time on which this one may be expected to have impact. Choice strictly in accordance with (U), then, is *myopic choice*. An agent chooses myopically by "treating the choice to be made at each point in the decision tree as if it were an isolated choice, unconnected not only with what came before but, even more important, with anything that can be projected about the choices he will subsequently make [RDC 12; also see PRR 219]." An only slightly less myopic agent might make such projections and assume he will follow through on his projected future choices when the moment occurs, but be regularly and predictably wrong.

Myopic choice expresses a *separability* condition on rational choice, i.e. that given a temporal sequence of choices within an action plan, any choice indexed to a particular point within that sequence is approached as though it were the first within the action plan, i.e. as though the particular branch of the decision tree at which the choice point appears were the origin of the tree. At each such point, the consequences of previous choices are regarded not as prior commitments with which one must coordinate one's present choice, but rather as external environmental constraints on present choice to which one bears no deliberative relation. Like other external events, they impose merely causal restrictions that condition the background against which the choice is made.

> What is decided at time t has no force at time t+1, unless at t+1 there is independent ratification of that plan from the consequentialist [i.e. utility-maximizing] perspective of t+1. That simply implies that the notion of a commitment to a plan has no meaning in the context of [separability] [RDC 208].

It is as if one proceeded from one choice to the next rather like Clyde in Volume I, Chapter IV.3, who rethinks all of his priorities from one moment to the next, erasing past choices from his memory as he turns his attention to the next one.

Myopic choices often engender what I called in Volume I, Chapter IV.3 time-dependent psychological inconsistencies and what McClennen, Strotz, and Hammond call *dynamic inconsistency*.[3] Dynamic inconsistencies occur when later choices contradict or subvert the intended consequences of earlier choices. The paradigmatic example is that of Ulysses' later rebellion, upon hearing the Sirens' song, against his earlier resolve to ignore them and continue on his way home. McClennen's example is of resolving to diet in the morning, then violating that resolve at that evening's dinner. Strotz finds an empirical generalization in such examples. He says,

> An individual is imagined to choose a plan of consumption for a future period of time so as to maximize the utility of the plan as evaluated at the present moment. His choice is, of course, subject to a budget constraint. Our problem arises when we ask: If he is free to reconsider his plan at later dates, will he abide by it or disobey it – *even though his original expectations of future desires and means of consumption are verified?* Our answer is that the optimal plan of the present moment is generally one which will *not* be obeyed, or that the individual's future behavior will be inconsistent with his optimal plan.[4]

The challenge this raises for McClennen's project is that such time-dependent inconsistencies in preference imply that the agent repeatedly thwarts her attempt to maximize utility over time: She prefers at t_1 to diet at t_3, but prefers at t_3 not to observe the diet she chose at t_1 for t_3. Thus the choice intended to maximize future utility is later thwarted by a contradictory choice that undermines it. Multiply this pattern over many instances as Strotz suggests, and the unhappy conclusion is that agents generally ignore or subvert rule-guided behavior that maximizes utility. The simple meta-rule, to act as we have resolved to act, appears not to guide our behavior even when it would maximize utility to do so. So in order to defend rule-guided behavior that maximizes utility, McClennen must find a psychologically viable alternative to myopic choice.

3. Precommitment and Sophisticated Choice

Strotz mentions two such alternatives. The first is precommitment, which Strotz describes as "precluding future options so that it will conform to his present desire as to what it should be."[5] Ulysses' strategy of having his men tie him securely to the mast so as to prevent him from following the Sirens would be an example of precommitment. So would some of the examples Strotz offers, of joining the army, getting married, savings plans and insurance policies whose low rates of return are, contra Strotz, justifiable as the price of ensuring adherence to the plan chosen at t_1. For the delinquent dieter, precommitment might most radically involve arranging with a dentist to have her jaw wired shut.

McClennen's objection to precommitment is two-fold. First, it is alienating; it limits an agent's freedom. And second, it seems irrational. It requires expending scarce resources on the

project of imposing inviolable and involuntary constraints on one's future behavior. An agent who chooses precommitment

> ends up expending resources that do not have to be expended [by simply following the diet], resources that are valued both from the standpoint of the time of planning, and the time at which the plan is to be executed. Fees must be paid to join a diet club, extra effort must be expended to keep the wrong kinds of food out of reach, or one must risk the disapproval of one's friends, etc. On the assumption that you continue to prefer more to less money, prefer not to risk the ridicule of your friends, etc., what you do … is to create a real intrapersonal dilemma for yourself. In effect, the only "rational" sequence of choices you can make leads to an outcome that can be characterized as intrapersonally *suboptimal*, since both from the perspective of the time of planning, and from the perspective of the time of execution of the plan, you disprefer that outcome to the outcome of [simply following the diet]. [PRR 234]

McClennen's analysis also implies that precommitment is fully consistent with myopic choice, since it at t_1 imposes involuntary constraints on choice at t_3 that at t_3 become merely part of the environment of natural events and states of affairs to which the agent's preferences at that moment must respond. In precommitment, all the work of enforcing at t_3 the agent's choice at t_1 is done by the external constraints because the agent is not assumed at t_3 to bear any deliberative relation to the choice he made at t_1:

> The agent who precommits "ties the hands" of his future self; that is, he "deposits his will" in some external structure, so that when he arrives at the subsequent choice point, certain options are no longer available [RDC 158].[6]

An agent whose hands are thus tied is compelled by those external constraints to perform within a certain restricted range of actions at t_3, and so need not connect that performance with any previous choices made at t_1 – or, for that matter, any future ones to be made at t_n.

Strotz's second alternative to myopic choice is what McClennen calls *sophisticated choice*. Here the agent chooses an action plan based on informed projections of how she will choose later in reaction to commitments made earlier; and rejects now those that she knows she would in any case choose to reject later. "To be sophisticated, then is to tailor your *ex ante* choice of a plan to your projection of what you would prefer, and hence choose, *ex post*." [PRR 221] McClennen regards sophisticated choice as the more inclusive concept of which precommitment is an instance, because both involve choosing a plan based on projections of future behavior. However, precommitment effectively forecloses projected *ex post* choices, whereas sophisticated choice adapts to them.

On the face of it, sophisticated choice seems to be an alternative to myopic choice because it involves consideration at the outset of the entire sequence of choices constitutive of an action plan. But it is not a true alternative, because part of what the agent considers and incorporates into the final plan is the likelihood that she later will reject certain options myopically. So she now sophisticatedly rejects plans containing choice point options that she knows she later will

myopically reject. But this strongly suggests that the plan she now sophisticatedly chooses consists in choice point options she later myopically chooses.

For instance, take the dieting case. Suppose Audrey knows now that she later will be unable to stick to the diet she is now considering, so she now decides not to embark on the diet in the first place. But if she knows she later would myopically reject dieting if she now chose it, why should she assume that her later choice to continue not dieting is any less myopic, or is as considered as her choice now not to do so? Just as she later would have reconsidered the choice to diet had she made that choice now, she later may similarly reconsider her choice not to diet assuming she makes that choice now. Then her decision tree looks this way:

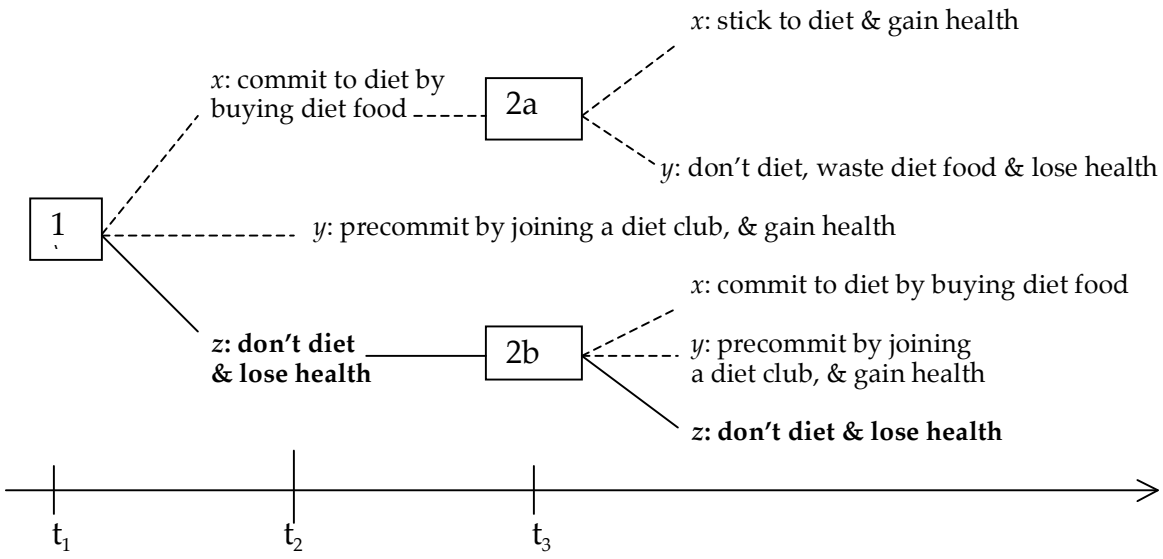


Figure 3. The Sophisticated Myopic

That Audrey made at t_1 a sophisticated choice not to diet does not commit her to carrying through with that choice at t_3, even though it incorporated Audrey's knowledge that she at t_3 would not follow a diet had she chosen it. There is nothing in the notion of choice adapted to projections of future behavior that requires such future behavior to be deliberatively sensitive to past choices. Then Audrey's choice of *z* at t_3's choice node 2b bears no less a myopic relation to her choice of *z* at t_1's choice node 1 than her choice of *y* at t_3's choice node 2a would have borne to her choice of *x* at t_1's choice node 1. Since in neither case is the relation between her choice at t_3 and her previous choice at t_1 a deliberative one, the consistency of her choice of *z* both at t_1 and at t_3 is fortuitous.[7] McClennen observes that the separability that this form of myopia expresses is for many inextricably linked with utility-maximization itself [RDC 207], and that sophisticated choice is fully consistent with separability [PRR 230].

4. Resolute Choice

Both precommitment and sophisticated choice are strategies by which the agent adopts action plans that enable him to circumvent the exercise of will, whether through the imposition of external constraints or the avoidance of choices that require it. Since there is no deliberative connection between present and future choices, no exercise of will is needed to carry it through. McClennen's solution is *resolute choice*, which includes both. He proposes that "[r]ather than regimenting present choice of a plan to projected future choice, the required alignment can be secured, in principle, in just the reverse manner, by *regimenting future choice to the originally adopted plan* [PRR 231; italics in text]." Absent new information or changed circumstances, a resolute chooser carries out his commitment in the future to the plan he selected earlier. That plan imposes an internal constraint of will parallel to the external constraint imposed by precommitment. It is an implicitly sophisticated plan to the extent that the agent rejects commitments he knows he lacks the will power to keep, e.g. a commitment that requires him to violate rationality conditions to which he is committed [RDC 201], or that requires him to conform to rationality conditions that are too demanding [RDC 202].

However, McClennen offers a further incentive[8] for a resolute chooser to conform his future choices to his originally chosen plan – an incentive in terms of which, he argues, strength of will may be understood: that resolute choice is less costly than any of the alternatives. Homer's resolute choice to diet looks this way:

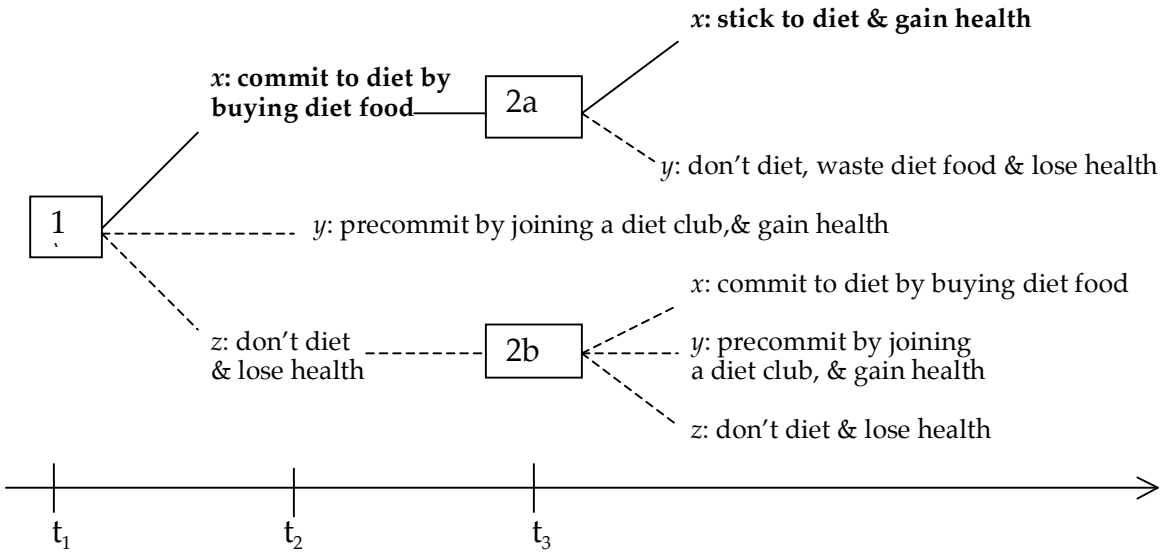


Figure 4. The Resolute Chooser

By coordinating t_3's choice node 2a with his decision at t_1 to diet, Homer avoids wasting his diet food, damaging his health, and expending time and money on a diet club. The savings are considerable.[9] Are they greater than if Homer had attempted simply to impose tyrannically the

preferences he has at t_1 on his future choices? Yes. For unless Homer understands the "tyranny" of choosing at t_3 in a manner preferred at t_1 but not at t_3 to be offset by the overall savings a resolute chooser gains, Homer fails to maximize utility at t_3, incurring a loss that might well outweigh the utility of imposing his preferences at t_1 on the future. The difference between tyranny and resolution is deliberative comprehension. Deliberative comprehensiveness sacrifices separability in order to cure myopia. Thus one aspect of resolute choice for McClennen involves more than merely imposing one's will on future preferences that conflict with it. It involves shaping future preferences in light of present rational deliberation as to how future choices may maximize overall utility when coordinated with present ones:

> Choice within the decision tree is shaped by a plan that is responsive to the totality of prospects that he confronted at the outset. For such an agent, choice points within the decision tree are continuation points: He sees his task (at each such point) as that of continuing to implement the plan he initially settled upon, so as to ensure that the sequence of choices thus made serves to access the prospect he initially judged to be most acceptable (or, at the very least, took to be one of those that were acceptable) [RDC 158-59]. … [t]he ex post resolute self is oriented to the idea of the ex ante self as a controlling self and, hence, to the idea of his ex post self not being completely independent. RDC 160]

In this case present, utility-maximizing rational deliberation engenders a rule that guides and regulates future choices, and shapes the agent's preferences accordingly.

A second dimension of resolute choice speaks to the question of how the ex post self can be brought to observe the resolve the ex ante self makes, given that its preference ordering contradicts that of the ex ante self. McClennen's answer is that the ex post self has an incentive to observe the ex ante self's resolve under those circumstances in which resolute choice is the optimal outcome intrapersonally for both ex ante and ex post selves. Suppose, for example, that Irene at t_1 wishes to diet at t_3, knows that at t_3 she will abandon that plan, and therefore is disposed at t_1 to have her dentist wire her jaw shut instead. Knowing at t_1 that at t_3 she will prefer to respect her wish at t_1 to diet at t_3 rather than have her jaw wired shut, even though her first choice at t_3 would be to abandon her diet, Irene at t_1 can use the threatened alternative of having her jaw wired shut to motivate herself at t_3 to stick to the diet she chose at t_1. If she knows at t_3 that her predicted abandonment of the diet she chose at t_1 disposed her at t_1 to implement the more discomfiting alternative of having her jaw wired shut, then she knows at t_3 that she is getting off easy by sticking to her diet. Gratitude and relief can be powerful incentives. Thus Irene at t_1 and Irene at t_3 can both agree that sticking to her diet serves Irene's interests at both times: her interest at t_1 in not abandoning her diet at t_3, and her interest at t_3 in not having her jaw wired shut at t_1. Sticking to her diet is a solution to the problem of coordinating the conflicting interests Irene has at each of these two times:

	stick to diet	wire jaw shut	abandon diet
Irene at t_1	5	4	1
Irene at t_3	3	1	5

Figure 5. An Intrapersonally Coordinated Resolute Chooser

Thus in intrapersonal cases that have this kind of Prisoner's Dilemma choice structure, resolute choice can be spelled out as solving a coordination problem between the conflicting interests the self has at different times. Again the incentive for being resolute – for ensuring consistency between the choice made at t_1 and the action taken at t_3 – is the awareness that violating one's resolve is costly both from the earlier and from the later standpoint. McClennen's proposal is one that argues

> for a model in which the plan that is taken to be regulative of subsequent choice is one that can be defended from the perspectives both of the time of planning and the time of choice. That defense turns on the consideration that the kind of coordination over time that planning makes possible economizes on scarce resources that are valued both at the time of planning and the time of choice. [PRR 241]

McClennen's concept of resolute choice offers a single agent a two-fold utility-maximizing justification of rule-guided behavior under some circumstances: First, it is justified when both earlier and later selves see the cost of violating the earlier resolve. Second, it is justified when both earlier and later selves prefer it to the costs of intrapersonal conflict.

5. Resolute Choice and Genuine Preference

Actually McClennen's concept of resolute choice is justified even when neither of these conditions obtain. That is, it is justified even when utility in the unreconstructed, minimalist sense of (U) is not maximized. This means that it can be, after all, "unhinge[d] from what [McClennen] take[s] to be its basis, namely, pragmatic considerations [RDC 160, fn. 12 (285)]." McClennen himself resists this conclusion. However, I argue in Section 7 below that thus unhinging resolute choice from questions of utility-maximization has a consequence McClennen endorses, namely it not only leaves open but in fact implies

> the possibility that even if the agent did not as a matter of fact resolve at some point before n_i to choose in a certain fashion at n_i, still one can consider as relevant to the question of what is to be chosen at n_i what one would have resolved to do at some antecedent point if one had (counterfactually) considered the matter [ibid.].

That is, decoupling McClennen's model of resolute choice from its utility-maximizing preconditions exposes its nomological character and identifies it as not merely a rule but rather what Kant would describe as a law of rationality.

The dilemma of naïve (as opposed to sophisticated) myopic choice represented in Figure 5 and described in Section 2 above can be expressed in the terms used in Volume I, Chapter IV.2 to describe a cyclical ranking (C_t):

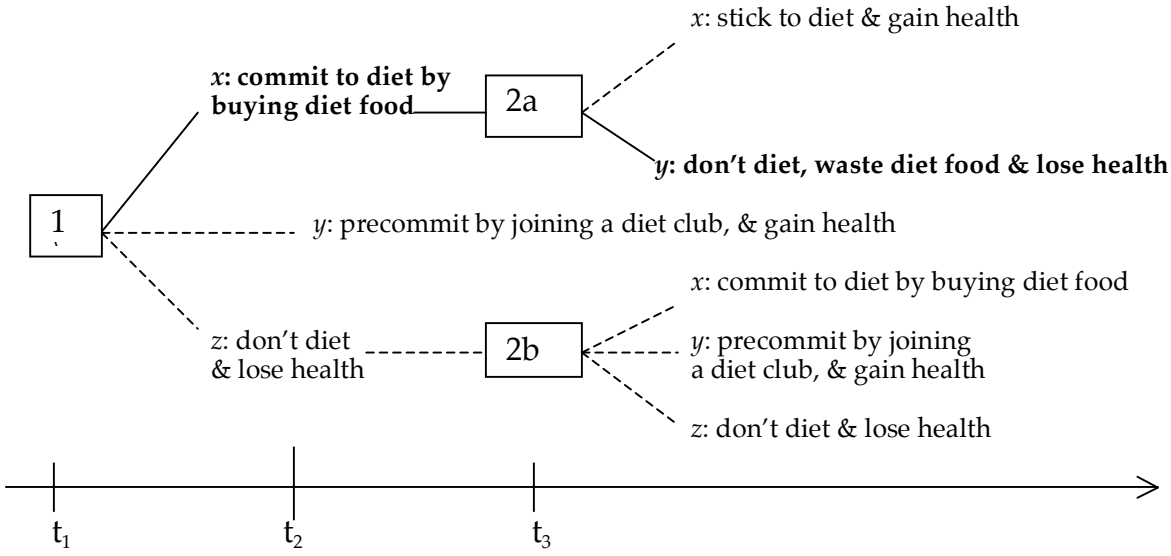


Figure 6. The Naïve Myopic

At t_1 Dennis prefers that at t_3 he voluntarily follow his diet (F) over having his jaws wired shut (G); at t_2 he prefers having his jaws wired shut (G) to breaking his diet at t_3 (H); and at t_3 he prefers breaking his diet at t_3 (H) to voluntarily following his diet at t_3 (F):

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

Dennis' preferences – and therefore his actions – are cyclical over time, and the problem to which McClennen's model of resolute choice is a solution is the problem of cyclical choice examined at length in Volume I, Chapter IV.2 – 3 [cf. RDC 89 – 98]. Dennis chooses, as McClennen observes, "as if he had blinders on – as if he never considered anything but the immediate choice problem presented to him at each point in time [RDC 97; also 206 – 209]." Correspondingly, the rule violated by Dennis' myopic choice behavior is the rule of transitivity; and the rule-guided behavior McClennen defends is transitively consistent behavior.[10]

In the notation proposed in the preceding chapter, (C_t), becomes Chapter III.9's

(C') *Pw(x.~y).Pw(y.~z).Pw(z.~x)*

There I argued that the solution to the problem of cyclical choice is to require that an agent such as Dennis choose only on the basis of his genuine preferences. To act on one's genuine preferences is to act resolutely in McClennen's sense: to preserve the transitivity, among other things, of one's preference orderings. In Volume I, Chapter IV.2 – 3, I defended the claim that

preserving transitivity is not the same as maximizing utility in the nonvacuous sense. In this volume's Chapter III and again here I extend this claim further: transitively consistent behavior can be justified even when it does not maximize utility in the nonvacuous sense.

The requirement that one act only on one's genuine preferences speaks to both of the utility-maximizing situations McClennen targets as justifying resolute choice. Consider the first. Suppose utility is not maximized when the later self follows through on the earlier self's resolve. Suppose instead that there is a considerable cost to so doing: Phoebe resolves at t_1 to drive her sick friend Timothy to the hospital at t_3, but at t_3 is inclined to choose the overall less costly alternative of paying a limousine service to do so instead, even though nothing in the situation has changed and there is no new information. Is there any other reason for Phoebe to nevertheless follow through on her original resolve despite its cost? The concept of a genuine preference provides one. As we have seen in Chapter III, the very fact that acting on her original resolves maintains the horizontal and vertical consistency of Phoebe's experience over time and at each moment is itself a reason. That is, preserving a unified and internally coherent self is a good that justifies Phoebe's resolve even though that unified self fails to maximize utility on this occasion.

Now McClennen describes and rightly dismisses a superficially similar case, in which one simply "might be the sort of person who values choosing in a manner that is consistent with earlier choices made [PRR 239, fn. 44];" someone who "simply ha[s] a preference for acting subject to the constraints of such rules [PRR 215]." In this case preserving transitivity through time would maximize utility. However, McClennen is right to reject this possibility as *ad hoc*, since whether one has such a preference or not will depend on arbitrary and idiosyncratic factors that do not require any special kind of motivation of the sort resoluteness provides. My claim, that preserving the internal unity and coherence of the self over time and at each moment justifies resolute choice independently of utility-maximization, is a different one. My claim is not about a contingent psychological preference for consistency, but rather about the necessary metaphysical consistency that genuine preference – indeed, any kind of preference – presupposes: I may or may not have a particular liking for consistency; but unless I am a unified and internally consistent self in the first place, the issue of my psychological likes and dislikes cannot arise. In Chapter V.2 below I argue that although preserving an internally coherent self in this sense is a good, it is not an end, goal or intentional object that an agent can adopt or at which he can aim. Therefore while it can be a justifying reason for action, it cannot be the object of a preference. So my claim is not susceptible to McClennen's criticism.

Next consider the second utility-maximizing situation that justifies resolute choice in McClennen's account, in which the later self follows through on the earlier self's resolve in order to avoid what it regards as an aversive outcome, and what the earlier self regards as only a second-best alternative; i.e. in which resolute choice coordinates an optimal outcome for both earlier and later selves. Suppose the earlier self does *not* have an inclination to choose the second-best outcome in anticipation of the later self's abandonment of the original plan: Myron at

t_1 rationally anticipates that Myron at t_3 will abandon the diet he choose at t_1, yet unlike Irene, cannot bring himself to adapt to this by having his jaw wired shut. So Myron at t_3 need not accommodate Myron-at-t_1's decision to reject this most aversive – for Myron at t_3 – outcome. Nevertheless, Myron at t_3 still has a reason to abide by Myron-at-t_1's resolve to diet. The reason is that his resolve at t_1 makes coherent and intelligible his sticking to his diet at t_3, whereas it makes abandoning his diet at t_3 incoherent and disorienting: Sitting stuffed, queasy and stupefied at his dinner table after having gorged himself on food he had for good reason resolved to forego, Myron is naturally confounded by the empty plates and distended expanse of stomach before him. He asks himself, Did he really eat all that? And why? To where did the sober and disciplined person he was at t_1 disappear? Myron chooses to avoid this condition of disconnected bewilderment by sticking to his diet at t_3, quite aside from the threat of precommitment or cost in resources of abandoning it. Once again, acting on a genuine preference is itself a good that, by ensuring the internal unity and coherence of his self at each moment and through time, justifies Myron's resolve.

6. Two Psychologies of Choice

McClennen approaches the psychology of rational choice differently than I do. Whereas in Sections III.2 and III.8 of the preceding chapter I spoke of an agent's ability to form and apply over time the concept of a thing's ranking superiority, and of her memory of previous pairwise comparisons, McClennen speaks of an agent's commitment to regulating future choice by an originally adopted plan; and of coordinating present choice with earlier choice. McClennen's psychology of choice is superior to mine in two important respects which I discuss in Sections 7 and 8, below. But overall, I (unsurprisingly) prefer my psychology of choice to his – not least of all because I think it accords better with his avowal that weakness of will can be understood as a "sign of imperfect rationality [PRR 236]."

McClennen's concepts of commitment, regulation, and coordination are "thick," psychologically complex ones that presuppose my more basic, "thin" psychological notions of evaluation and memory. That is, in order to now carry out a commitment to act on a plan earlier adopted, or to coordinate a present choice with a previous one, an agent must be able to form and apply consistently through time the concept of the ranking superiority of the earlier plan, and so the concept of the ranking inferiority of the present, cyclical preference on which she is now disposed to act. Additionally, the agent must be able to remember the relation of the two alternatives she is presently ranking – the original plan to the cyclical one – to the third she is not – the threatened alternative of precommitment.

Now we saw in Section 1 that McClennen cashes out the psychologically complex notion of commitment in terms of a more basic psychological disposition to follow the rule. But even a psychological disposition presupposes my yet more basic elements of evaluation and memory. In order to be overridingly disposed to do x rather than y or z, an agent must evaluate doing x as superior to doing y or z. He must also be able to form and apply consistently through time the

concept of *x*'s ranking superiority, and so the concept of *y*'s and *z*'s relative ranking inferiority. He must be able to remember the relation of *x* to *y* and *z* from – at the very least – the moment before the disposition is activated to – at the very least – the moment it is actualized. And if it is a real *disposition* to so behave, he must be able to do this not just on one occasion, but repeatedly, whenever the disposition is prompted.

But satisfaction of my psychological requirements is not only a necessary condition of McClennen's. If the argument of Chapter III is valid, it is a sufficient condition as well: If an agent satisfies the conditions of concept-formation and application described in Chapter III.8.(a) and (b), then he is effectively disposed to coordinate his later choice with his earlier one as the concept of a genuine preference requires – and as McClennen's conception of rule-guided behavior requires. As McClennen rightly suggests, resolute choice requires an exercise not of will, resolve or commitment in the ordinary sense; but rather of reason. Reason is exercised when alternatives are consistently ranked and the consistency of that ranking through time is maintained, whether so doing maximizes utility or not; this is another example of the sense in which utility-maximization in the nonvacuous sense is a special case of, but not co-extensive with the more comprehensive, Kantian conception of rationality elaborated in Chapters II and III. Formally, McClennen's model of resolute choice is different, and much more technically sophisticated than my concept of a genuine preference. But psychologically, they are materially equivalent; and McClennen's concept of resolute choice can be justified by the considerations I offer in defense of genuine preference, independently of whether or not deliberatively regulating later choices in light of earlier commitments maximizes utility. I put both McClennen's and my psychologies of choice to work in dissecting the character of the whistle-blower in Chapter VI.8, below.

Whether utility is maximized or not, rational agents have not only a justification but also an incentive to conform their behavior to McClennen's model of resolute choice. We saw in Volume I, Chapter IV.2 that failing to choose consistently over time undermines the very possibility of unified agency. I extended that argument in Chapter II.6 above, and do so even further in Chapter V.2 – 3 below. In this context, McClennen's own witty characterization of the picker's dilemma is worth quoting again in full:

> How is [Clyde] 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. [RDC 208]

McClennen's description of the picker's dilemma applies *a fortiori* to that of the chooser's. We have also seen in Chapter II.4 – 6 above that failing to organize coherently all of the experiences constitutive of one's perspective at a particular moment similarly subverts the capacity for unified agency. Among one's experiences at a particular moment are memories of earlier choices and concepts consistently applied. Then an agent's incentive for deliberatively regulating later choices in light of earlier ones is the same as the justification for so doing: to preserve the internal unity, consistency and coherence of the self both at each moment and through time. I defend this claim at greater length in Chapter V.2 – 3. An agent has this incentive whether utility is maximized or not.

If a rational agent has both justification and incentive for acting resolutely – i.e. on her genuine preferences – independent of utility-maximizing considerations, then the commitment – i.e. the psychological disposition – that guides the behavior of a rational agent is not dependent for its activation on the deliberative conclusion as to whether or not utility is thereby maximized. That psychological disposition is continually active, prompted by the continuing pressure of incoming sensory data from both the external and the internal environment, and by the agent's own bodily and psychological responses to them. It functions by continually incorporating and organizing information, experiences, and the resulting preferences at each moment in relation to previous ones. I represent this disposition as part of a decision tree in Chapter V.2 (Figure 7) below.

7. Nomologicality and Kant's Derivation of Promise-Keeping

To say that the disposition to resolute choice is continually active is to say that it is not merely a rule that guides utility-maximizing action. It is a categorical indicative *law*, in Kant's sense, that in the ideal case of perfect rationality guides all behavior, both action and impulse, both actual and possible. I further develop this concept of a law in Chapter V.5.2 below. As a law, resolute choice satisfies the nomological requirement that it apply universally in both actual and counterfactual cases: If an agent chooses *x* at t_1, she will honor that choice at t_3; if she were to choose *y* at t_1, she would honor that choice at t_3; and if she had chosen *z* at t_1 she would have honored that choice at t_3. So McClennen's concept of resolute choice enjoins a rational agent to canvas each of the possible choice alternatives available at each moment with an eye to whether she can carry through in the future on the plan of action her present choice implies; that is, to make no choice that is deliberatively disconnected from past or future ones. Similarly, because resolute choice decoupled from utility-maximizing considerations functions as a law-like criterion of rational choice, it evaluates each choice with an eye to its deliberative consistency with earlier choices, whether or not that choice was in fact was the outcome of an earlier

resolution. In effect it requires of each choice that it be such that it *could have been* the outcome of an earlier resolution.

For example, reconsider Phoebe's deliberation at t_3 as to whether she should pay a limousine to take her sick friend Timothy to the hospital, or drive him herself. Is the first option consistent with earlier choices she would have made regarding Timothy's well-being if she'd thought about it? Probably not: the difference between a paid limo and Phoebe's car is Phoebe's concerned and reassuring presence, the palpable support of a friend in Timothy's time of need. If Phoebe really is Timothy's friend and not merely a patron, well-wisher or well-meaning bystander, then the impersonality of merely paying for his trip in a limousine would remind Timothy of Phoebe's palpable absence, a reminder that is inconsistent with such support. So Phoebe at t_3 must now choose as though she is carrying through on an earlier commitment to drive Timothy to the hospital, even if she earlier made no such commitment. For the even earlier commitment she did make was to be Timothy's friend. Resolute choice requires all of her behavior toward Timothy thereafter to be consistent with that. More generally, resolute choice requires that for any choice of *x* at t_n, *x* satisfy the consistency criteria for a genuine preference listed in Chapter III relative to earlier relevant choices ("relevance" being defined by Chapter III.9's (VC) and (VC^P)), even if the agent did not explicitly resolve at t_{n-m} to so choose at t_n. McClennen's model functions as a criterion of rationality relative to which all preferences, whether utility-maximizing or not, are evaluated. Barring changes in circumstances or additional information, it enjoins the law-governed consistency in choice that a genuine preference requires.

Above I offered some reasons for preferring my psychology of choice to McClennen's. So I did not mention two considerations in terms of which McClennen's is superior, when suitably decoupled from the issue of utility-maximization. The first is that the thick concepts of commitment, resolve, regulation, and coordination that undergird McClennen's psychology of choice identify the model of resolute choice as the general law of which the rational necessity of promise-keeping is a special case. If a rational agent later honors earlier choice commitments, then in particular a rational agent later honors earlier choice commitments uttered performatively to another agent. That is, a resolute chooser by definition keeps his promises. Now promise-keeping is a special case of resolute choice in that it invites a more elaborate and complex social justification than that offered here. But the fact that keeping one's promises in particular is implied by choosing resolutely in general would claim a foundational role in any such justification. Thus despite his explicit resistance, McClennen's model does even more than "develop a 'deontic' theory of resolute choice that would form the analogue to, say, a theory of morality in which the fact of having promised was taken as sufficient to establish an obligation [RDC 160, fn. 12 (285)]." Resolute choice is not merely an analogue for such a theory of morality. It is at its foundation.

Kant's argument for promise-keeping in Chapter I of the *Groundwork* offers a contrasting strategy of justification. He offers as a general criterion of rationality the concept of "bare conformity to law as such (without laying at its basis any law determined by particular

actions[11])" which, he says, "serves the will as its principle." [G, Ak. 402] I argue elsewhere that this criterion is merely a summary reformulation of the criteria of rationality Kant develops at length in *The Critique of Pure Reason*. He believes that from this criterion it is possible to derive certain specific moral obligations, promise-keeping among them. He applies this "principle of the will" to the maxim of "getting out of a predicament by a false promise [G, Ak. 402]." He asks whether this maxim could "be valid as a universal law (one valid both for myself and others) [G, Ak. 402]," and concludes that "because it cannot fit as a principle into a possible enactment of universal law [G, Ak. 403]," false promising is to be rejected.

Kant's reasoning here does not conclusively dispose of false promising, however. It is possible to tinker with the formulation of this or any maxim in such a way that a suitably revised version *could* "fit as a principle into a possible enactment of universal law." For example, the above maxim might be specified in greater detail as "getting out of the difficulty of being robbed at gunpoint by a false promise that the check written to the robber for one's checking account balance will not bounce." It is hard to see how this maxim might fail to "fit as a principle into a possible enactment of universal law." On the other hand, reformulating the maxim so as to include the agent's knowledge that the robber has a starving family to feed and cannot find work, and that the checking account in question contains the smallest balance of many the agent has might rather exclude false promising under these circumstances from becoming universal law. Similarly, it is hard to see why the maxim of keeping a promise to make executor of one's estate someone who turns out to be a confidence man should fit into such an enactment. There are as many examples pro and con as there are act-descriptions.

The same difficulty infects Kant's derivation of all the specific moral practices he considers: truth-telling, preserving one's life, helping others, and cultivating one's talents. Generally speaking, Kant's problem is that simple universalization of a maxim is by itself too weak a criterion to rule out all cases of false promising, lying, suicide, self-neglect, or any other practices that are, in their most general descriptions, *prima facie* morally unacceptable. His rationalist disdain for "any law determined by particular actions" leads him to adopt as law a principle so weak and comprehensive in scope that it potentially justifies virtually all actions given sufficient specification of the circumstances. It is not fine-grained enough to distinguish between those specific actions that really are justifiable and those which are not.

McClennen's model of resolute choice avoids this problem because, as a pragmatist, he has no objection to laws "determined by particular actions," provided that the particular action prescribed is broad enough in scope so as to include a sufficiently diverse range of instantiations. Thus McClennen's pragmatic approach makes a virtue of the generality that is a defect in Kant's rationalistic approach. The particular action McClennen prescribes is the resolute regulation of later choices by earlier ones. As a law, resolute choice is instantiated by sailing past the Sirens, sticking to one's diet, staying married, not deserting the army, and many other actions. It is also, more generally instantiated by keeping promises we make, whether to ourselves or to others – and so by any contract we make. However, because it is based in the "particular actions" and

choices that Kant abjures, resolute choice does not require keeping promises or honoring contracts independently of the informed preferences and known circumstances on which such choices are based. Thus McClennen's model of resolute choice succeeds where Kant failed in the *Groundwork*, namely in deriving a more fine-grained and suitably qualified conception of promise-keeping from the very concept of reason.

8. Free Riding and Moral Emotion

Because McClennen's model of resolute choice can be justified independently of utility-maximization, it provides an even more "secure footing for a *rational* commitment to practice rules [PRR 215; italics in text]" than McClennen himself claims, and so an even more fertile solution to the free rider problem. As we saw in Volume I, Chapter XII, the free rider is an agent who enjoys the benefits of others' compliance with a rule but breaks it when this is personally advantageous. Tax evasion, welfare fraud, insurance fraud, accounting fraud and failure to contribute to public radio would be examples. If everyone reasoned as the free rider does, there would be no benefits to enjoy. Because the free rider's reasoning is equally available to everyone, free riding is a threat to the continued existence of those benefits. So the challenge is to somehow discourage free riding – either by showing the reasoning to be defective, or by establishing viable social sanctions against it.

One reason why the problem has seemed intractable to some philosophers is that it has been viewed as a strictly interpersonal coordination problem.[12] Hobbes' original introduction of the problem begged the question of how to solve it by stipulating that the Fool "*declares* he think it reason to deceive those that help him, … [he] breaketh his covenant, *and consequently declareth* that he thinks he may with reason do so,"[13] as though in acting to obtain the personal advantages of breaking a rule that others keep, the free rider thereby announced to those others his violation of it. Under these circumstances the distinctly less than clever free rider of course would have reason to heed Hobbes' warning as to the dangers of getting caught (Fool that he is). But this would make free riding nothing more than a pointless exercise in self-destructive behavior. Hobbes offered no viable answer as to why a clever, and therefore secretive free rider should not exploit for personal gain others' compliance with the rule, because there is no room in his proto-Humean conception of the self for a noninstrumental, supervisory role for reason. For Hobbes, if considerations of personal advantage justify entering into the social contract, considerations of personal advantage similarly justify breaking it under certain circumstances. That is, Hobbes' Fool lacks a conscience.

Kant's answer to the free rider is similarly unsatisfactory, for several reasons. First, even if the free rider had a conscience of the sort that functioned in the manner of Kant's noninstrumental conception of reason, we have seen that it would still be possible to justify violating many beneficial social covenants simply by tinkering with the formulation of the maxim. Second, Kant's principle of "bare conformity to law as such" requires only that the free rider entertain the counterfactual conditional of whether the rule violation could be

universalized. Even if the relevant maxim could not be reworked so as to satisfy this requirement, attempting to universalize it would demonstrate only the social and political impossibility that everyone could behave as the free rider does in fact. It would not demonstrate that this particular free rider should not so behave, assuming others do not do so as well. Indeed, the very conceptual possibility of a free rider depends on the assumption that most other people do not behave similarly. Third, therefore, no such counterfactual appeal is likely to move the free rider to reform her ways, because her proto-Humean psychology is such that she lacks the moral and rational susceptibility to such an appeal.

> Finally, Mill's practical solution, that
>
> laws and social arrangements should place the happiness or (as, speaking practically, it may be called) the interest of every individual as nearly as possible in harmony with the interest of the whole; and, secondly, that education and opinion, which have so vast a power over human character, should so use that power as 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[14]

abandons the attempt to find flaws in the free rider's reasoning, and instead opts for a radical form of social coercion that simply eliminates it, along with the very ability to conceptualize self-interest altogether. Mill basically proposes that the resources of law, social sanction, and education should be deployed as tools of propaganda to reprogram all individuals, by erasing any concept of or motive to self-interest and reconditioning them to identify their interests with those of the community. In the wake of the fall of Communism we have sufficient empirical evidence to conclude that such social programs are inherently untenable. At each stage in their implementation, uncooperative free riders can always be found. What Hobbes', Kant's and Mill's solutions have in common is that they conceive the free rider problem as one that arises when individual utility-maximization conflicts with that of the community. Their solutions are unsuccessful because they all invoke the interests of the community to discourage free riding in an agent for whom the interests of the community are irrelevant.

McClennen's solution takes a different route. Resolute choice conceives the free rider problem as one that *first* arises when an individual's utility-maximization conflicts not with the community's, but rather with itself over time. If McClennen's model of resolute choice were dependent on intertemporal utility-maximization, it would provide only a similarly conditional solution to the free rider problem. It would discourage free riding only in those cases in which the intertemporal maximization of utility justified keeping the social covenant to follow shared rules, but not otherwise. Free riding then could be justified in any case in which the utility gained by breaking this promise outweighed the loss of intertemporal utility consequent on the broken promise itself. For example, while tax evasion would violate an agent's earlier

commitments and implicit obligations as a citizen, this loss of intertemporal utility might be outweighed by the gains of engaging in it. The cyclicity of his myopic choice might be a small price to pay for the lifestyle the tax evader would be able to enjoy.

Decoupled from utility-maximizing considerations, however, totaling up these gains and losses is unnecessary and irrelevant. Resolute choice discourages all cases of free riding because it invokes the value of intertemporal consistency itself – i.e. of maintaining one's individual agency *simpliciter* – to discourage free riding in an agent whose preservation of his own agency must take unconditional priority. No luxurious lifestyle would be worth the threat to unified agency and coherent self-determination that myopic choice expresses. What's the use of having a lot of money if you have to depend on other people to remind you what it's for and how you've spent it? Resolute choice suggests that the basic problem with free riding – the one that surfaces even before the free rider's exploitation of others' compliance – is that broken promises disconnect the free rider not only from others but, even more seriously, from himself.

Now let us return to a vivid example of such disconnection, in order to trace whence the internal sanction against promise-breaking arises. That McClennen's model of resolute choice in effect derives a prima facie obligation of promise-keeping from a rationality criterion explains why Myron (Section 5, above) would feel not only bewildered and disoriented by violating his diet, but also *betrayed* by his ungoverned impulses. His having resolved to diet in the first place evinces a self-conception that his later violation would disconfirm: He thought he knew who he was and what his capacities for self-governance were, but this violation would prove him wrong, weak, self-deceived, and fat.

From this self-betrayal would arise a mix of at least three central moral emotions: guilt, shame, and resentment. – Guilt, because Myron would have inflicted a harm by breaking a rationally and morally justified rule. The victim would have been himself, the harm would have been ill health, and the rule broken would have been that of resolute choice, to abide by one's commitments. Consequently Myron would view himself as morally derelict with regard to his own long-term well-being. Second, he would feel shame, first because of his multiple failures to live up to his own, idealized self-conception; and second because of the way his impulsive behavior at t_3 exposed his multiple personal and moral flaws to the disapproval and ridicule of his own, self-critical eye. Third, he would feel resentment toward his earlier self for misleading him as to his true capacities for self-governance; and toward his later self for demonstrating how minimal those capacities in fact were.

Notice that a myopic chooser does not suffer these painful emotions, because she does not recur to earlier choices when dealing with the unpleasant consequences of later ones. Hence although her later violation of an earlier preference ranking may lead her to feel just as queasy and disoriented as Myron, she does not feel self-betrayed by breaking the diet she earlier chose to keep, because that earlier choice bears no deliberative relation to the present one. Of course a myopic chooser bears no stronger a deliberative relation to the preferences of other people than she does to those of her earlier self.

The painful emotions of guilt, shame and resentment, consequent on breaking a promise made to oneself, provide additional incentive to Myron to stick to his diet. Myron is naturally disposed to avoid not only threats to his internal unity and coherence as an agent; but the negatively reinforcing self-dislike these emotions can cause. McClennen's psychology of resolute choice implies that these moral emotions, and the self-dislike they instill, can arise solely out of the self's relation to its own earlier incarnations, independently of certain community-wide norms, practices and values inculcated during the process of socialization.

Now recall that McClennen aimed to show that a psychological disposition to rule-guided behavior could arise from rational deliberation alone, independently of involuntary socialization or hard-wired biological drives. I am not convinced he has succeeded in demonstrating the complete independence of resolute choice from hard-wired biological drives, either on his utility-maximization interpretation or on my "deontic" interpretation, because the disposition to maximize utility or to preserve the internal unity of the self themselves may be biologically hard-wired. However, I do think he has shown both interpretations of resolute choice to be independent of "altruistic gene" accounts of biological hard wiring. And the foregoing sketch of the genesis of moral emotion suggests that McClennen has definitely shown resolute choice to be independent of involuntary socialization.

But because resolute choice implies promise-keeping, and the disposition to promise-keeping can be instilled, in part, by the aversive effects of the painful moral emotions consequent upon breaking promises to oneself, McClennen has shown more than this. It is not only resolute choice that is engendered by deliberative rationality independent of involuntary socialization. Human morality itself can be deliberatively engendered in exactly the same way. McClennen's model of resolute choice, suitably decoupled from his insistence on utility-maximizing considerations, implies that human morality is much more closely entwined with deliberative rationality than most Humeans would agree. A close look at McClennen's model of resolute choice reveals him to be – like so many in the lineage of American Pragmatism – at heart a Kantian without the metaphysics.

Endnotes to Chapter IV

[1]Peter Hammond, "Consequential Foundations for Expected Utility," *Theory and Decision 25* (1988), 25-78.

[2] Edward F. McClennen, *Rationality and Dynamic Choice: Foundational Explorations* (New York: Cambridge University Press, 1990) [henceforth RDC], 83; also 144-146. Also see his "Pragmatic Rationality and Rules," *Philosophy and Public Affairs 26*, 3 (Summer 1997) [henceforth PRR], 223 and fn. 22. Page references to both works are cited hereafter in the text.

[3]The concept of myopic choice was originally introduced in R. H. Strotz, "Myopia and Inconsistency in Dynamic Utility Maximization," *The Review of Economic Studies 23*, 3 (1955 – 1956), 165-180, and significantly developed in Peter Hammond, "Changing Tastes and Coherent Dynamic Choice," *The Review of Economic Studies 43* (1976), 159-73; and "Dynamic Restrictions on Metastatic Choice," *Economica 44* (1977), 337-50. Although my approach to the problem in Volume I, Chapter IV.2.3 and 3, as well as my solution in this chapter are quite different, they converge on the same issues that are under scrutiny here.

[4] *Ibid*. Strotz, 165.

[5]*Ibid*.

[6] This latter phrase was coined by Jon Elster. See his *Ulysses and the Sirens: Studies in Rationality and Irrationality* (New York: Cambridge University Press, 1979), 43.

[7] Audrey's choice of *z* at t_3's node 2b is more costly in the long term than any other option, and more costly at t_3 than it was at t_1. So from a utility-maximizing perspective it is not only myopic but irrational.

[8] I use the word advisedly. McClennen speaks of justification rather than incentive. But on a Humean conception of the self, justification and motivation rely on the same element, namely desire. A justification that does not appeal to desire can in theory do no motivational work, and a motivation that does not appeal to desire is by definition impossible.

[9] For the technical version of the argument, see RDC Sections 9.6 and 11.4 – 6.

[10] Because I grounded Chapter IV's analysis of genuine preference primarily in the strict preference relation for simplicity's sake, here I similarly confine my remarks to questions of transitive versus cyclical rankings, ignoring issues that arise out of the distinctions between transitivity and acyclicity, and between intransitivity and cyclicity. Optimistically, I assume my demonstration in Chapter III.6.2.1 above, that transitivity and acylicity are logically equivalent, to be dispositive of any such issues.

[11] Incredibly, Paton translates *<<ein auf gewisse Handlungen bestimmtes Gesetz>>* as "any law prescribing particular actions."

[12] Here I discuss only historical approaches to the problem. But for a useful overview of contemporary Humean approaches that reformulate the free rider problem as the problem of public goods and connects it with the Sorites paradox, see Richard Tuck, "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.

[13] Thomas Hobbes, *Leviathan*, Ed. Michael Oakeshott (New York: Macmillan/Collier Books, 1977), 115; italics added.

[14] John Stuart Mill, *Utilitarianism*, Ed. George Sher (Cambridge: Hackett Publishing Co., 1979), 17; italics added.

Chapter V. How Reason Causes Action

Having constructed the formal framework of transpersonal rationality established in Chapters II through IV, I begin in this Chapter to flesh it out with some of the richer psychological phenomena of rationality that accommodate its requirements. Specifically, I now offer an account of what Kant calls the causality of reason, i.e. the power of rational principles and considerations to motivate action in the ideal case. I shall say that *principles and considerations are rational* if they satisfy the constraints of the framework already established. If, as I have just argued in Chapter IV, McClennen's concept of resolute choice is materially equivalent to my concept of a genuine preference, and resolute choice provides an intrapersonal foundation for moral commitment, then it could be argued that the concept of a genuine preference in effect provides such a foundation, and therefore entails the relationships of trust and responsibility that a stable interpersonal morality must presuppose. However, I shall not attempt any such Deductivist argument here. My primary task is to show that rational principles, and in particular the rational principles constitutive of a genuine preference, can have the motivational efficacy that McClennen's idealized account takes for granted.

I consider two ways in which reason can have motivational efficacy: first as a necessary condition and contributing cause of action; and second as sufficient condition and precipitating cause of action other things equal. The first accounts for reason as a necessary condition of what I call *literal self-preservation*, i.e. the preservation of the internal unity and rational coherence of the self, according to the criteria of rationality proffered in Chapters II and III; and therefore as a necessary condition of action of any kind. This is the theme of Sections 1 and 2. Section 3 contrasts my account of reason as a necessary condition of action with Marcia Baron's analysis of duty as a secondary motive of action. Baron's analysis focuses on specifically moral motivation, whereas mine targets rational motivation in general, of which moral motivation is (at least for purposes of this project) merely an instance. But Baron's version of a Kantian account of motivation highlights some of the differences between a "New Kantian"[1] approach to the issue and the Ur-Kantian approach I take here.

The second way in which reason can have motivational efficacy is as a sufficient condition and precipitating cause of action other things equal. I address this case in Section 4. Section 4.1 argues that the burden of proof should not be on the Kantian to prove that cortically trackable, occurrent thought- or belief-events do have motivational efficacy, but rather on the Humean to prove that they do not. Section 4.2 argues, against Baron's repudiation of primary motives, that they are essential to Kant's account of moral motivation. Section 4.3 distinguishes between the power of an occurrent thought- or belief-event to precipitate action merely in virtue of the conative power of the event itself, and its heightened power to precipitate action in virtue of the conative power of its content as well. Section 4.4 focuses on this second case as the one in which the antecedent thought- or belief-content governs, directs and guides the consequent action whose own intentional content reflects it; i.e. in which action is determined by will. Here I

make a three-fold distinction between a motivationally ineffective intellect, an opportunistically effective intellect, and a motivationally effective intellect. This last-mentioned is the key to the concept of strength of will, and to the argument that reason can precipitate action. I distinguish three kinds of motivationally effective intellect: one for whom reason motivationally overrides conflicting inclination; a second for whom reason and inclination are each sufficient and conjointly overdetermine action; and a third – which I call a fully effective intellect – for whom reason is the only source of motivation there is. In Section 4.5 I focus on this third case, and argue that a fully effective intellect generates descriptive principles of rational agency that it believes to be true, and true of itself; and that an agent who implicitly recognizes herself in these principles is prompted by them to actualize her rule-governed disposition to rationality in action that instantiates them.

Section 5 applies this analysis of ideally rational motivation in general to ideally moral motivation in particular (without, however, committing to any Deductivist relationship between them); specifically, to Kant's concept of the perfectly good will as always and only motivated by reason. In Section 5.1 I extend Kant's account by way of an analogy with the motivational efficacy of a principle of nonmoral, logical reasoning on actual human behavior. In 5.2 I then analyze the actual principles of Kant's substantive moral theory. I argue that conjointly, they meet familiar criteria for being a genuine theory that describes and explains the behavior of a perfectly rational being; that these principles therefore qualify as theoretically rational according to the consistency requirements established in Chapters II and III; and that they therefore move to action an agent who implicitly recognizes herself in these principles.

Finally, in order to place in a broader context the concepts of the causality of reason by contrast to the causality of desire, I invoke in Section 6 a distinction between two ideals of rational motivation that finds its origin in Nietzsche: the ideal of spontaneity versus the ideal of interiority. This contrast sheds light on how my account of a fully effective intellect could have application to some actual agents under certain conditions; how, that is, reason could override desire to precipitate action in an actual agent. It also explains in greater depth why such an agent is accurately described as transpersonally rational. I argue that the ideal of spontaneity makes the concept of a motivationally effective intellect inexplicable, whereas the ideal of interiority makes it unremarkable.

1. Rational Action

Chapter III defined a genuine preference as one that satisfies not only the criteria of horizontal and vertical consistency introduced in Chapter II, but also additional consistency criteria most of which are familiar from classical logic: asymmetry, connectivity, irreflexivity, transitivity, and ordinality. We saw in Chapter II that the first two ensure that a genuine preference is rationally intelligible, i.e. it is recognizable as an instance of concepts that partially constitute an agent's perspective at a particular moment. We then saw in Chapter III that the additional five further ensure that a genuine preference preserves that logical consistency and

conceptual coherence over its entire duration, however long or short that is. Conjointly, these criteria represent a genuine preference as rational in virtue of the theoretical rationality of the concepts by which the agent represents that preference to himself. In Volume I, Chapter II I defended a representational theory of desire. Genuine preferences, then, include desires the agent's representations of which satisfy the requirements of theoretical rationality.

However, genuine preferences comprise more than desires in this modified Humean sense, for the reasons mentioned in Volume I, Chapter VI.4.2 and taken up in greater detail below: Agents can and do choose to pursue valued intentional objects, i.e. ends, which they nevertheless have no desire (in the nontrivial sense) to pursue. Any end an agent chooses that satisfies the above consistency criteria counts as a genuine preference, whether that end is the object of a desire – or, alternately, of a resolve or mere intention. I shall say that *an agent acts rationally* when all of the ends for which he acts satisfy these consistency criteria; and that *a rational action* is one whose particular ends do as well, regardless of the type of motive that moves him to act.

2. Literal Self-Preservation

Now we saw in Chapter II above that *all* of the concepts constitutive of an agent's perspective at a particular moment *must* satisfy the minimal criteria of theoretical rationality expressed in the requirements of horizontal and vertical consistency, in order for her experiences to be rationally intelligible to her. This of course includes her ends: in order for an agent's ends to be rationally intelligible to her, they must be genuine preferences in the sense just explained. But we also saw in Chapter II that the requirement of rational intelligibility is actually quite a compelling one, since an agent who violates it cannot make sense of her experience, nor therefore conceive of herself as having – nor therefore as authoring – her own experiences. Violation of the requirement of rational intelligibility thus doubly undermines the capacity for autonomous agency.

So if the promise of rational intelligibility is the carrot that disposes an agent to seek only those ends that satisfy the above consistency criteria, the threat of psychosis is the stick that discourages her deviation from them. A unified agent is disposed, above all, to act in ways that preserves the capacity for agency, for rational intelligibility, and ultimately for coherent selfhood. On this thesis, the rational unity of the self is preserved when all of the experiences that constitute the agent's perspective are rationally intelligible in the sense already explained. An agent is overridingly disposed to preserve the rational unity of her self when this disposition culls from her experience any objects or concepts, including ends, that violate the above consistency criteria. Such an agent is disposed to avoid actions, behaviors, or experiences that undermine the unity of the self, and to react aversively when they are forced upon her or compelled by causal determinants (whether inner or outer) outside her control. I shall express this by saying that a unified, rationally coherent human agent by definition has what I shall call a *highest-order disposition to literal self-preservation*. Essentially this is a streamlined version of Kant's

synthetic unity of apperception. My contribution to Kant's idea is to make explicit what Kant very clearly implies: that this unity is structured by principles of theoretical rationality – i.e. of logic, and that coherent agency would be impossible without it.

2.1. Motivational Efficacy

To be motivated by reason is at the very least to be moved overridingly by the highest-order disposition to literal self-preservation in the sense just described. This just is the preservation of the rational intelligibility of our experience in the form necessary for agency, i.e. as self-conscious experience. In Chapter II we also saw that this, in turn, requires that the ways we conceptualize our experiences satisfy at least the requirements of horizontal and vertical consistency, however else they may differ. These requirements, I argued, are the familiar requirements of theoretical reason applied to the substantive and predicative constituents of declarative categorical propositions we occurrently – but not always explicitly – believe. This means that the disposition to literal self-preservation is, in effect, preservation of theoretical rationality as motivationally overriding in the structure of the self. Theoretical rationality is motivationally overriding in that it constrains and is a necessary condition of any other motive an agent may have, including desire. For without it, there would be no coherent agent to be motivated to perform any particular action whatsoever. So literal self-preservation must be an essential, hard-wired disposition that any such action – and any more particular motivation for it, including desire – must presuppose. Literal self-preservation – and therefore theoretical rationality – enables us to preserve the horizontal and vertical consistency over time of the highest-order concept of our selves as having our experiences, and so constrains all other motives we as agents can have.

2.2. A Good But Not an End or a Desire

That literal self-preservation has survival value implies, of course, that it has value, i.e. that it is, for us, a normative good. But we have just seen that literal self-preservation just is the preservation of the rational intelligibility of one's experience, i.e. satisfaction of the requirements of horizontal and vertical consistency over time. This means that what we often refer to as descriptive or explanatory coherence is itself a normative good – one we must achieve to some degree before we can even attempt to achieve any other.

However, not every normative good can be adopted as an end. The descriptive or explanatory coherence of an agent's perspective is not itself an end a rationally unified agent can pursue – nor, therefore, a genuine preference an agent can have, because this coherence is a precondition for the formulation of any end an agent can pursue. This normative good is the valued outcome that secures the rational integrity of that perspective in the first place; a necessary condition any such preference or end must presuppose. Because the highest-order disposition to literal self-preservation constructs, governs and preserves the theoretically rational integrity of an agent's perspective, it itself can bear the self-consciousness property – i.e. of being

an experience the agent has – only under certain esoteric conditions that are irrelevant to this project. Hence it itself is not subject to the demands of horizontal and vertical consistency, nor to the five additional consistency criteria discussed in Chapter III. This highest-order disposition is instead what causes a nascent self to satisfy them. However, other preferences that are subject to these criteria must satisfy them relative to the enduring highest-order disposition to literal self-preservation. To appropriate McClennen's terminology, literal self-preservation (*LS-P*, below) is a permanent disposition that constitutes an *internal* "environmental constraint" relative to which a rationally coherent agent orders all such relatively transient genuine preferences *a*, *d*, and *g* through time:

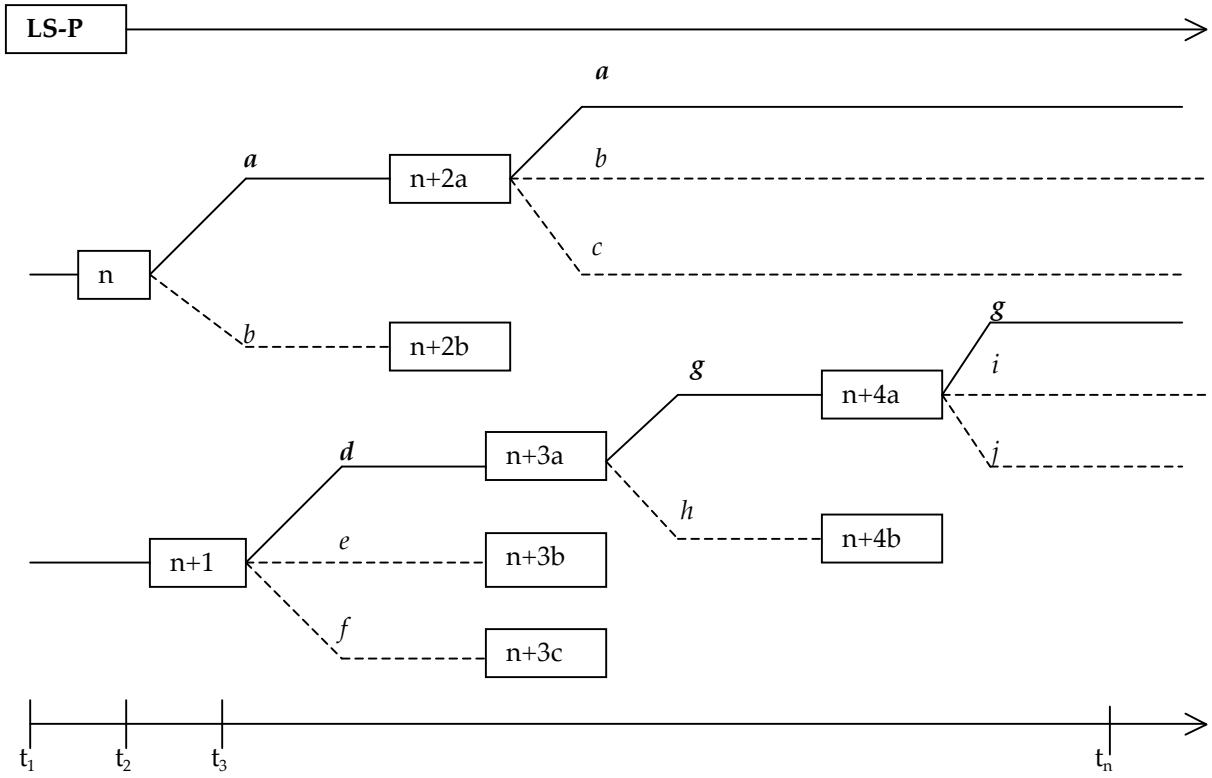


Figure 7. The Highest-Order Disposition to Literal Self-Preservation

In Figure 7, LS-P lacks a choice node origin because at no point is it the object of a choice; it is a necessary condition that any choice presupposes. Similarly, it does not lead to a choice node because it does not generate any *specific* choices; it imposes an environmental constraint that any such choice must satisfy. Hence the relation between this highest-order disposition and an agent's genuine preferences is asymmetrical. Although LS-P need satisfy no further consistency criteria relative to them, they must satisfy the above consistency criteria relative to it. The highest-order disposition to literal self-preservation provides the enduring backbone against which the consistency of an agent's preferences – and so his personal continuity – is measured. Therefore it is the metaphysical backbone against which the mere psychological preference for

consistency itself, which McClennen described in Section 5 of the preceding chapter, must be measured.

That the alternative to preserving the theoretically rational integrity of the self is some degree of psychological disequilibrium also explains why this disposition to satisfy the consistency criteria of theoretical rationality could not be redescribed as a desire to satisfy these criteria. For an occurrent desire to do or pursue something (the only kind, as we saw in Volume I, Chapter II that is worthy of the name) is an experience a unified and rationally coherent agent has. So it cannot be the same as that disposition which is required by the continuing existence of the agent to begin with. We have seen that a moving, sentient envelop of flesh padded by layers of muscle and fat and draping a rack of bones cannot be a unified human agent unless at least minimal theoretical rationality criteria of horizontal and vertical consistency over time are met. The disposition to satisfy them is a necessary condition of unified agency. So in order for an agent to have a desire of any sort, satisfaction of these criteria – hence a disposition to literal self-preservation – must be presupposed. The disposition to literal self-preservation must be presupposed by any desire an agent has because it must be presupposed by motivationally effective agency. If it is a necessary presupposition of desire, it cannot be identical to desire.

For the same reason, it would be a mistake to suppose that the preservation of the theoretically rational intelligibility of the self might be a mere means to the satisfaction of some further desire. The highest-order disposition to literal self-preservation does have at least instrumental value, since it is a necessary precondition of any of the ends an agent adopts, and so *a fortiori* of those he actively tries to achieve at a given moment. But since this highest-order disposition also rules out the adoption of any ends that are themselves inconsistent with the rest of his experiences, it similarly rules out any such ends that might seem to have intrinsic value.

For example, consider the protagonist of Henry James' "The Last of the Valerii," a young Roman count of ancient lineage who unearths a pagan statue on his family's estate. The statue evokes in him the desire to engage in ancient and, to him, ultimately inexplicable Dionysian rituals of sacrificial worship. He finds himself compelled to perform these sacred rites nightly from dusk till dawn. Tormented by impulses that, although harmless, are to him ultimately unintelligible and inconsistent with the other desires and habits that characterize him as a modern European, he allows his wife to rebury the statue, rather than utilize his wealth and freedom to indulge these anomalous impulses. The fundamental disposition to literal self-preservation requires the suppression not only of external but of internal events that violate the requirements of horizontal and vertical consistency over time, on pain of cognitive and conative paralysis – or, at worst, madness.

2.3. Pain and Physical Self-Preservation

The highest-order disposition to literal self-preservation can be invoked to explain our aversion to physical injury, pain, or assaults of any kind, either physical or psychological; and adds context to the representational analysis of aversion offered in Volume I, Chapter II.2.2. I

argued there that an aversion was a complex emotion including not only distaste, revulsion or oversensitivity to a representational object, but also quite distinctive and visceral feelings of sensory overload, oversatiation, discomfort, and anxiety. Physical pain is the limiting case of violence to – i.e. violation of – the rational integrity of the self or ego through maximally acute, vivid, and intense sensory overload. But the same brand of violence, in more adulterated forms, is effected by any such violation – physical or psychological – of the requirements for the rational intelligibility of its experience. Pain is inherently disruptive of the ordinary processing of experience, even if we are vigilant, or have been forewarned; it is upsetting, confusing, and destabilizing of our psychological equilibrium at least for that reason. Its occurrence is cognitively and emotionally disorienting, and its persistence effective in restructuring an agent's entire personality and perspective to accommodate it.[2] Moreover, though we can explain its causes in terms of other events, we usually do not try to make pain rationally intelligible in any more fundamental terms than the concept of pain itself: We think of it as a basic kind of experience. But this may be mistaken. On the above hypothesis, pain is merely the most extreme case in which the integrity and boundaries of the self are violated by intense sensory overload. This would be the higher-order kind that in fact identifies it.

A corollary of this analysis of pain, then, is that the highest-order disposition to literal self-preservation explains the desire for self-preservation in the purely physicalistic, Hobbesian sense. If the body constitutes the outermost physical boundaries of the self (rather than merely a physical container in which the self is housed), then physical self-preservation is preservation of the rational integrity of the self against a particular kind of assault, namely bodily injury or deprivation. Bodily assault counts as an assault only because of the pain that ordinarily accompanies it. However, if pain itself is aversive only because it maximally overloads, disrupts and disorients the self and subverts its capacity for agency, then physical self-preservation is of value to us only as a means to literal self-preservation. If we did not care to defend the rational integrity of the self against such disorientation and subversion, the physical integrity of the body would be of considerably less concern to us (as those who have experienced the peculiar effects of laughing gas at the dentist's, or of minor surgery under local anaesthetic may agree). On this view, then, literal self-preservation, and the avoidance of self-disintegration, are biologically and psychologically more fundamental than physical self-preservation and the avoidance of pain as such. That is why physical self-preservation can be the object of a desire or preference whereas literal self-preservation cannot.

3. Baron on Secondary Motives

That literal self-preservation cannot be the object of a desire or preference differentiates it from what Marcia Baron calls a secondary motive. Baron's concern is to explicate what Kant plausibly might have meant by saying that an agent with a good will is motivated to act from duty rather than inclination. She aims to show that Kant's stipulation is satisfied in case duty is a secondary rather than a primary motive for such an agent. Adapting a distinction made by

Barbara Herman,[3] Baron proposes that duty operates as a *primary motive* "if it is the main impetus, the thing that moves me to act."[4] I explore Baron's conception of a primary motive further in Section 4.2 below. But Baron wants to show that the most important and illuminating account of moral motivation is to be found in the notion of a *secondary motive*.

A secondary motive has three defining features on Baron's view. First, it acts as a *limiting side-constraint* when other motives actually prompt one to act (113, 129). In this capacity it is not only possible but necessary that other motives provide the sufficient condition of action; the secondary motive merely endorses one's plan (130-131), and filters out impermissible maxims of action (144). These passages clearly differentiate such constraints provided by duty as a secondary motive from the primary motives thus endorsed or rejected.

This first feature of a secondary motive is similar to my account of reason as a necessary condition of coherent agency, i.e. literal self-preservation, in two respects. First, my account specifies certain consistency criteria that any rationally intelligible motive – indeed, any rationally intelligible experience more generally – must meet. These criteria thereby function as limiting side-constraints on experience that implicitly endorse some as rationally intelligible and exclude others as unintelligible. Second, these consistency criteria themselves cannot be regarded causally efficacious events, or "thing[s] that move me to act." These consistency criteria are abstract propositional objects that denote a certain metaphysical and structural condition of the agent whose experience and behavior satisfy them. As such, they furnish no sufficient psychological content – i.e. no intentional objects, no purposes, goals or ends an agent might adopt or by which an agent might be inspired to act. By themselves alone, they furnish structured place-holders for internally consistent experiential content. They are not the kind of entities that might motivate anyone to do any one thing in particular.

But my consistency requirements of literal self-preservation are unlike Baron's conception of duty as a secondary motive, in that these requirements must first be met in order for an agent to adopt any such purposes, goals or ends as genuine preferences. By contrast, on Baron's conception, duty as a secondary motive is not a necessary condition of coherent experience or action in general. An agent can act coherently in the absence of this limiting condition, i.e. can violate the constraints of duty without violating the constraints of rational intelligibility. Whereas rational intelligibility is a necessary condition, duty as a secondary motive is a contingent condition of unified agency. I think Baron's conception of duty departs significantly from Kant's in this respect, but I defer discussion of that point to another occasion.

A second characteristic of duty as a secondary motive is that it involves making a conscious and deliberate *commitment* to regulating the agent's conduct in accordance with what is right (113, 129, 140 fn. 22). This does not mean that he ignores or devalues the inclinational motives that prompt him to act (131), nor that he never fails to act according to the requirements of duty:

> [A] "perfect record" in doing one's duty is not only not sufficient to acting from duty but also not necessary. … one can correctly be said to act from duty even if one occasionally

> fails to do what one sees one should do. But the commitment will have other manifestations besides conformity to one's sense of duty, most notably, reflection on how one ought to live, readiness to revise one's moral beliefs and one's plans and aims in light of one's reflections, and willingness to entertain evidence that tells against one's moral beliefs. …The sense in which one acts from this commitment, even in instances in which [he] gives no thought to the ethical nature of [his] conduct before proceeding with the intended action, is roughly as follows: a very rich explanation of any nontrivial choice or action, e.g. the sort of explanation that a novelist might give, would make reference to some of the manifestations listed above. (140 fn. 22)

Thus a commitment to acting from duty as a secondary motive involves self-evaluation, introspection, reflection, and receptivity to rethinking one's beliefs and priorities in light of evidence and experience. It is a conscious choice about how to fashion one's life over the long term, including one's affective motives and sentiments, even if one fails to meet this standard in one's behavior on a particular occasion.

This set of characteristics of a secondary motive differentiates it from the consistency constraints of literal self-preservation on two counts. First, as observed earlier, literal self-preservation cannot be an end or desire, even though it is a normative good, because it is a necessary condition for an agent's having ends or desires. Therefore it cannot be the object of a conscious and deliberate commitment. Of course one can choose to make such a commitment to cultivate the virtue of consistency – for example, by keeping a journal in order to supplement one's memory of past actions and events; or by interrogating one's daily choices with reference to choices one has made in the past; or by keeping a planner, or taking notes on professional activities, and the like. These practices would involve the same kind of self-evaluation, introspection, etc. that characterize duty as a secondary motive on Baron's view. But in order to make such a commitment to the virtue of consistency, the more fundamental consistency requirements of literal self-preservation already would have to have been met. The conscious commitment to consistency might certainly strengthen and enhance this basic highest-order disposition, as well as articulate it consciously as an explicit value. But only a unified self is capable of such a commitment, which thus builds self-awareness and self-determination on a prior foundation of literal self-preservation.

Second, on Baron's conception of duty as a secondary motive, the long-term preservation of this motive is compatible with occasional violations of duty, because the commitment involved is to a certain form of long-term conduct, including complex practices of self-analysis and deliberation, in addition to action. Long-term literal self-preservation is also compatible with occasional violations of rational intelligibility, but for a different reason. By contrast with Baron's secondary motive of duty, long-term literal self-preservation is not ensured by supplementary long-term commitments and mental practices relative to which a single delinquent action or experience has little import. Rather, it is ensured by the repetitive, selective functioning of those consistency requirements themselves: The consistency requirements of literal self-preservation

impose criteria for the selection of each and every experience that is rationally intelligible to the agent whose experience it is. So experiences, motives and actions that violate them fail to achieve rational intelligibility, thus fail to be integrated into the agent's perspective, and so fail – at least in the ideal case – to violate the rational integrity of the self. That is, functioning consistency criteria of literal self-preservation *protect* the rational integrity of the self against external threats to its unity. Only when too many such experiences violate these criteria does the rational intelligibility of the agent's perspective begin to fail; and only then does the preservation of the unity of the self come under threat. I examine this case at greater length in Part II below.

This second set of characteristics of the commitment to duty as a form of long-term conduct imply a third, that duty as a secondary motive is *not token-guiding*. An agent need not be constantly preoccupied by her duty, nor with trying to ascertain what her duty is (131):

> A responsible moral agent should take an active interest in a wide range of moral questions…. None of this says … how often she should reflect on these things. What is important is that … she should be committed to changing herself … or less isolated in her own affairs, and more aware of social injustices and prepared to contribute to ameliorating them. (132) … duty operating principally as a secondary motive … attaches primarily not to individual actions but to conduct, to how one lives, and only derivatively to isolated actions. It serves generally as a limiting condition and at the same time as an impetus to think about one's conduct, to appraise one's goals, to be conscious of oneself as a self-determining being, and sometimes to give one the strength one needs to do what one sees one really should do. ... [it plays the role of] prompting us to reflect on our conduct and in maintaining or heightening our moral sensitivity. (134)

Thus duty as a secondary motive does not require repeated, continual, moment-to-moment acts of conscious attention to each in the sequence of individual act-tokens an agent performs. It does not require the agent to monitor and evaluate each such act with respect to its accordance with or violation of the mandates of duty. This is the corollary of the second feature of duty as a secondary motive, that the long-term maintenance of this secondary motive is compatible with occasional violations of its strictures because it involves attention to long-term conduct rather than to "isolated actions."

In this regard, secondary motives are comparable to the consistency constraints of literal self-preservation. The former *need* not be enduring intentional objects within an agent's perspective; whereas the latter, as mentioned above, *cannot* be. However, the two are dissimilar in that in Baron's secondary motive of duty, one's attention is weighted toward long-term conduct *rather than* isolated act-tokens; whereas observation of the consistency constraints of literal self-preservation shape long-term conduct *in virtue of* screening each isolated act-token. In the ideal case considered here, individual actions that violate these constraints are not performed, whereas those candidates for action that satisfy them are. Over time, this selective mechanism through which consistent actions are filtered functions to habituate the agent to act consistently,

thus forging enhanced psychological support for the highest-order disposition to literal self-preservation that already is deeply inscribed in the structure of the self.

However, the passage directly above raises further questions. We have already seen that Baron means to distinguish acting from duty as a primary motive from so acting as a secondary motive; and I have collated only those passages in which Baron addresses acting from duty solely as a secondary motive. However, at page 134 above she first explicitly identifies this as her subject, but then describes duty as a secondary motive as serving "generally as a limiting condition and *at the same time as an impetus* to think about one's conduct … [and] *prompting* us to reflect on our conduct etc." [italics added] Here she uses the language of primary motivation – of, as she later describes it, "a force within us that causes us to act accordingly" (189); of "a picture of agency on which agents act from inner pushes or urges or tugs or drives" (191).

Now Baron means to repudiate such a picture on Kant's behalf. I think this is misguided, both on Kant's behalf and on her own; I air this opinion at greater length in Section 4.2 below. The points to be made here are, first, that Baron's own description of secondary motives is in fact compatible with such a picture, since presumably the process of fashioning a valued course of long-term conduct for oneself must proceed via moment-to-moment habituation, i.e. by occurrently reminding oneself of that value on at least most of the occasions of "isolated action" that are relevant to it. Otherwise the agent's commitment to acting from duty would not come to much. Second, therefore, Baron's conception of duty as a secondary motive is not only compatible with such a picture but also requires it, since unless there is some identifiable juncture at which "to think about one's conduct, to appraise one's goals, to be conscious of oneself as a self-determining being" actually translates into causal efficacy, it is very hard to see how such reflective mental activity could, in fact, "give one the strength one needs to do what one sees one really should do;" i.e. how it could be anything more than self-indulgent wheel-spinning.

However, if these two aspects of duty as a secondary motive are, indeed, not only compatible but necessarily interconnected, then duty as a secondary motive bears the same complex relation to isolated act-tokens on the conscious psychological level that the consistency constraints of literal self-preservation bear to isolated candidates for rationally intelligible experiences more generally on the pre-conscious metaphysical level. For in both cases, the process of fashioning the relevant behavioral dispositions is additive and cumulative. On the conscious psychological level, the agent deliberately undertakes a program of moral self-improvement that consists in cultivating certain attitudes and dispositions through habituation. Habituation consists in practicing the valued actions – i.e. in instantiating the relevant normative principles in individual act-tokens – on at least most of the occasions on which one is offered the opportunity. Once these principles are firmly embedded as attitudinal and behavioral routines, the agent's conduct over the long term will reflexively reinforce and extend them, cumulatively, with each such act-token that further instantiates them. Analogously, on the pre-conscious metaphysical level, literal self-preservation requires observation of the consistency constraints from moment to moment, even though the particular experiential candidates for rational

intelligibility have unforeseeable durations and may be replaced or altered over time. The highest-order disposition to literal self-preservation is similarly reinforced and extended with each occasion on which experiential content is successfully screened for inclusion in or exclusion from the agent's perspective. Over the long term, in the ideal case, this additive and cumulative process strengthens and deepens the rational integrity of the self.

4. Rationality as a Sufficient Condition of Action

So far I have described the highest-order disposition to literal self-preservation as a kind of sentinel that repels all such threats to the theoretically rational unity of the self, filtering out inconsistent or conceptually anomalous beliefs, desires and impulses, and admitting in only those that qualify as genuine preferences. To the extent that this account is correct, it provides an analysis of reason as a necessary condition and contributing cause of action. But can reason also be a sufficient condition and precipitating cause of action other things equal? Can reason itself – i.e. rational content that satisfies the constraints already discussed – incite an agent to do something?

I argued in Volume I, Chapter VI that in practice, agents can and sometimes do pursue ends which they have neither desire nor impulse to realize. And in Volume I, Chapter VII I argued that Thomas Nagel would have done well to defend the commonsense thesis that an occurrent belief is a psychological and neural event that can cause an agent to act when neither desire nor impulse is present. However, I also argued there that this much would have been insufficient to demonstrate the motivational efficacy of reason itself, because we as yet had no means for distinguishing between the motivational efficacy of the belief-event qua event, and that of its belief-content. This leaves me with two tasks: first, to take on the thesis I reproached Nagel for avoiding, i.e. that an occurrent belief can precipitate action; and second, to show how the content of such a belief can direct and guide the action the belief-event precipitated. That the *rational* content of such a belief can do so will then follow straightforwardly.

4.1. How Thoughts Cause Action

By an *occurrent thought* or *belief*, I mean a psychologically discrete mental event or state that can be tracked cortically by way of a neurally discrete brain event or state. The content of an occurrent thought or belief need not be explicit, and therefore need not be the object of a conscious intentional attitude in order for the occurrent thought or belief to be a psychologically discrete mental event. I cannot give you an example of an implicit but occurrent thought or belief that is not the object of a conscious intentional attitude without turning it into one. But anything you now occurrently think or believe that you are not at this moment considering will do.

Also, there are many other kinds of belief besides occurrent thoughts and beliefs, not all of which are comfortably susceptible to a dispositional analysis. An example would be a pervasive, gnawing belief about your cosmic insignificance that saturates all of your reactions,

but to which you nevertheless would not be disposed to assent. For present purposes I leave all such thoughts and beliefs aside.

The belief-desire model of motivation claims that only desires can motivate action. As we saw in Volume I, Chapter VI.1, to accept this thesis is by definition to reject the possibility that other psychological states of the agent, such as thoughts or beliefs, might motivate action. It treats desire-causation as a matter of fact, and thought- or belief-causation as an unsubstantiated hypothesis. It thus enables Humeans to displace onto their Kantian opponents the burden of proof that occurrent thoughts, beliefs, or deliberations also can be motivationally effective independent of desire.

As we also saw in Volume I, Chapter XV, the rationale for the Humean position may be sought in an epiphenomenalist view of the mind, according to which mental contents are a nonmaterial and so causally impotent by-product of physical processes. However, this is not a convincing rationale without additional argument that shows occurrent desires to be exclusively physical events rather than mental contents as well, and occurrent beliefs to be exclusively mental contents rather than physical events as well. I doubt this can be shown. I doubt it is possible to demonstrate that any occurrent psychological event – whether thought, belief or desire – is not also a neural event; and am quite certain it is impossible to demonstrate that any neural event is not also a physical event with at least some degree of causal efficacy under some circumstances. I comment further on this rationale in Section 6.1, below.

Therefore I reject the Humean rationale for displacing onto Kantians the burden of proof of the motivational efficacy of occurrent thoughts and beliefs. Consequently I also reject that burden of proof. On the contrary: I claim that the burden of proof is on the Humean to explain how an occurrent thought or belief could be a psychological event without also being a neural and hence physical event; or if it is (even more improbably), how it could be a physical event that lacks any degree of causal efficacy under any circumstances. More specifically, I claim that the burden of proof is on the Humean to explain how these particular psychological events could constitute a breach in the causal network of empirical events, many of which nevertheless seem in retrospect to reflect or express the content of those antecedent psychological events. I challenge the Humean to explain how the constant conjunction of a psychological event and a subsequent behavioral event that expresses its content could be coincidental. Awaiting such an explanation, I assume in the meantime that occurrent thoughts and beliefs are psychological and neural events that take their places in the causal network of events just like any others. I take my task here to be to propose an account of how the causality of a certain kind of occurrent thought or belief might operate.

4.2. Baron on Primary Motives

My conception of an occurrent thought or belief would conform to Baron's definition, above, of a primary motive as "the main impetus, the thing that moves me to act." (113) Baron has two objections to a Kantian account of moral motivation that situates primary motives at the

center of such an account. First, she argues that the very idea of a primary motive belongs to an empiricist – i.e. a Humean – sensibility, and so fits poorly with Kant's conception of moral motivation. Second, she argues that Kant does not really need primary motives to explain how duty works. I disagree on both counts. Since moral motivation is the important instance of rational motivation that I ultimately aim to address, it will be convenient to dispose of these objections to the notion of an occurrent thought or belief as the key to moral motivation, before proceeding to explain the sense in which it is the key.

First let us clarify further what a primary motive is and how it functions. Baron says, "[M]y sense of duty may prompt me to refrain from doing something that I recognize to be wrong but am tempted to do, for example, to lie to save face." (129) In this case my sense of duty is a felt, consciously occurring psychological event that thwarts and overrides my temptation to lie to save face. It conforms to the model of "a force within us that causes us to act accordingly," (189) in this case to refrain from lying. However, the notion of acting according to a force deserves further scrutiny. The wind is a force that may push me across the street whether or not I am ready to cross it. But I do not act according to this force, for two reasons. First, I do not act at all; I am rather swept across the street. But suppose this were describable as an action, such that I intended to cross the street anyway and construed the force of the wind as helping me do so. It still would be peculiar to claim that I acted *according* to this force, as though the force issued directives to which my behavior conformed. I can act according to or in conformity only with something I interpret as providing a model I may or may not emulate, a template I may or may not fit, or directive that I may or may not follow. That is, I must ascribe to such a force some intentional content that is capable of guiding the behavior I undertake to perform. The wind has no such intentional content; one's sense of duty clearly does. The primary motive that prompts me to refrain from lying to save face, then, is an occurrent psychological event whose intentional content prohibits me from lying to save face. I would identify this event as an occurrent belief that I am not to lie in order to save face. I would claim that this occurrent belief is causally efficacious in thwarting and overriding my occurrent temptation to lie to save face.

However, Baron contends that that this is at odds with Kant's picture of agency. Her criticism falls into two parts: first, she objects to construing Kant's account of acting from duty in terms of motives at all:

> [T]he term 'motive' suggests a force that moves one to act, and yet the Kantian picture of agency is *not* one of agents being moved, but rather of agents acting for reasons or (to put it as Kant does) on maxims (134) … The problem is that the term 'motive' suggests causation, as if the motive of duty or a desire to help another were a force within us that causes us to act accordingly. (189) … 'motive' does indeed suggest that the agent is moved, yet on Kant's picture of agency the agent is not moved. So the difficulty is one of trying to capture a Kantian notion of acting from duty without suggesting that the agent acts from an inner 'force' … (191)

In these passages Baron contends that the correct account of Kantian motivation involves reasons, not causes. We are not moved to do anything and do not act from any "inner force." On Baron's reading, our sense of duty does not cause us to do anything; it provides us with reasons for doing something, or – to use Kant's term, maxims on which we do something.

Baron here rejects the assumption that Kantian agents are moved to do anything, that they are caused to do anything. She contends that they instead have reasons for doing things. This purported conflict between reasons and causes has a long history that I examined at some length in Volume I, Chapters VI and VII. Briefly, it stems from the Humean assumption that only desires can motivate action; and from the externalist inference from this that therefore only desires can be both reasons for and causes of action – from which it would follow that any reason for action that is not a desire is therefore not a cause of action. The notion that a reason for action that is not a cause of action might provide a viable Kantian account of rational agency is given further credence by an interpretation of Kant's account of freedom according to which an empirical act-token motivated by respect for the moral law is not caused at all.

This is not the place to enumerate the ways in which I believe such an interpretation of the texts to be misguided. For present purposes it might suffice, first, to note that in rejecting the assumption that Kantian agents are caused to act, this interpretation a fortiori rejects the assumption that they are caused to act by reasons that are not desires; and second, to note the many passages in which Kant takes for granted the assumption that agent *are* caused to act by reasons that are not desires. Following are a few from the first *Critique*, *Groundwork* and second *Critique* in which Kant speaks of the causality (*Kausalität*) of reason, freedom, or the will. *The Critique of Pure Reason*: 1C, A 317/B 374, A 328/B 385, A 444/B 472, A 446/B 474, A 534/B 562, A 537/B 565 and generally throughout the Resolution of the Third Antinomy; *Groundwork of the Metaphysic of Morals*: G, Ak. 446, 450, 452, 453, 457, 458, 460, 461, 462; *Critique of Practical Reason*: 2C, Ak. 6, 15, 16, 20, 21, 32, 42, 44, 45, 47, 48, 49, 50, 55, 56, 65, 67, 69, 70, 73, 75, 78, 81, 89, 94, 98, 103, 105, 113, 115. This last text is also rife with talk of the moral law as an incentive and as directly determining the will.[5] One might be able to make a plausible case that some of these passages should be ignored. It would be harder to make the case that most or all of them should be; and even harder to explain them all away. In these passages and many others, Kant assumes virtually without argument that reasons that are not desires can be causes.

But suppose Kant's assumption here is wrong. If rational agents on a New Kantian view are not moved to do anything, not even by their rational beliefs, how is it that they move into action at all? How is the transformation from static subjecthood to dynamic agency effected? The second part of Baron's criticisms aims to answer these questions by sketching what she takes to be the correct interpretation of Kant's account of agency. This, she says, contrasts with the 'motive' reading. This brings us to Baron's second argument, that Kant does not really need primary motives to explain how duty works:

> Kant's theory of agency is very different. Our actions are not the result of a desire or some other incentive that impels us. An incentive can move us to act only if we let it.

> (189) … The more appropriate Kantian focus is on conduct, viewed over a stretch of time and guided by reasons. Maxims, unlike motives, have no closer tie to individual actions than to courses of conduct; in fact, maxims connect more naturally to courses of conduct than to individual actions. (190) … agents must affirm the urge or push if it is to determine them to act accordingly, and … the sense of duty has a regulative function rather than merely impelling or prompting. (191) … on Kant's view we act on maxims, not from motives. (192)

On Baron's reading of Kant, causes of action can move us only if we "affirm" them; this is a gloss on Allison's incorporation thesis, that incentives can motivate only by being reflectively incorporated into the maxims of action.[6] Baron thus shares with other New Kantians the interpretation of Kant according to which rational moral actions are the result of conscious acts of reasoning and deliberation that resolve into explicit mental acts of affirmation of, incorporation of, or taking up of desires into universalizable maxims of action.[7]

Of course merely being affirmed by, taken up, or incorporated into an agent's universalizable maxim could not be *sufficient* for distinguishing the moral motive from non-moral ones that equally satisfied O'Neill's contradiction in conception test.[8] Consider the maxim,

> (1) From self-interest I make it a permanent rule always to keep my promises, in order to avoid even the possibility of social sanction.

– this would be what Baron calls a secondary motive; or the maxim,

> (2) Out of craving I verbally deny my addiction to gumdrops, so as to maximize my access to them.

– this would exemplify a primary motive on Baron's view. (1) licenses keeping promises for reasons of self-interest. (2) licenses lying for reasons of desire-satisfaction. On the face of it, both (1) and (2) formulate intentions that are universalizable without contradiction. Clearly, neither maxim furnishes a moral motive. What makes a motive a moral one is not merely its incorporation into a universalizable maxim. It must be the right *kind* of motive, which neither self-interest nor desire can be for Kant. Not even respect for the universalizability of (1) or (2) can be the right kind of motive on Kant's view. I offer an account of the right kind of motive in Section 5.1 below.

I do not agree with Allison's or Baron's account of maxims,[9] and, as indicated in the introduction to Chapter II above, I do not read Kant as requiring any such explicit and complex, conscious deliberative process as a prerequisite for morally worthy actions. Such a process makes moral agency seem too much like a rickety, 1950s physical simulacrum of a Turing machine, with grinding wheels, rotating sprockets and tape heads loudly whirring, that bears little relationship to the streamlined intuitive and pre-conscious processes which characterize our actual mental lives, and which I try here to rationally reconstruct.

However, Kant does often speak of adopting incentives into maxims, and I do agree that on Kant's view, in order for a potential cause of action to have motivational efficacy for an agent, she must occurrently conceptualize and thus recognize it consistently with the network of

concepts constitutive of her perspective as an agent. In this sense she must "incorporate" or *integrate* it into her internally consistent conceptual scheme. On my Ur-Kantian analysis, an agent integrates a motive into her perspective by recognizing that motive as an experience she has, and thereby conceptualizing it as an intentional object. If the motive is also an end or goal (and not all motives are), then she conceptualizes it as the object of her intention, and her maxim is her description of that intention. An intention such that its object both has causal efficacy and satisfies the consistency criteria of a genuine preference is both a cause of and a reason for action. I say more about other motives that are neither ends nor goals, yet also can be both causes and reasons, as well as enter into maxims and genuine preferences, below in Chapter VI.7.

By contrast, Baron is committed to denying the causal efficacy of any such "affirmations" or "incorporations" themselves. Yet why they should be supposed not to have causal efficacy is a mystery. If an agent's affirmation of the urge or push, rather than the urge or push itself, is what enables it to determine action, then surely the affirmation is functioning to boost the horsepower of that urge or push, and is therefore a primary motive in the same sense, with added horsepower (here I perversely revel in the provocation and shock value to New Kantians of the term "horsepower"). Similarly, Baron's account leaves no room for a causal explanation as to what might lead us to "affirm the urge or push [that] determine[s us] to act accordingly;" i.e. what causes us to occurrently conceptualize and recognize the motive as we do. There must be some explanation of why we affirm the moral law, or duty, as a causal determinant of our action; it cannot be that we do so for no reason at all.

Kant's answer is that it is the rationality itself of the moral law that occurrently causes us to do this. It is in this that the *Kausalität von Vernunft* consists, and for this that we feel *Achtung*. My Ur-Kantian account proposes that what motivates us to conceptualize the moral law as an object of respect and therefore a rational causal determinant of action is our occurrent recognition of the rational intelligibility within our perspective of a certain kind of maxim, which satisfies the consistency constraints definitive of a genuine preference. Some other kinds of morally controversial maxim will fail to satisfy some of those constraints. (1) above, for example, will violate vertical consistency within most agent perspectives because, as we have already seen in Volume I, Chapter XII and in Chapter IV directly above, permanent conformity to a rule of promise-keeping notoriously conflicts with self-interest rather than instantiating it; this demonstrates one sense in which self-interest is the wrong kind of motive to be incorporated into a specifically moral maxim. (2) above, by contrast, fails horizontal consistency because it is internally self-contradictory regardless of its motive: it both verbally asserts and verbally denies my addiction to gumdrops.

Hence both (1) and (2) exemplify maxims that are universalizable on the one hand, yet on the other, fail O'Neill's first part of the contradiction in conception test, that one "intend without contradiction to act on the maxim."[10]. Communities in which everyone attempts to promote but sometimes thwarts self-interest by always keeping their promises, or feeds an addiction to gumdrops by denying it are certainly possible. So if I can intend that any maxim of mine hold

universally,[11] I can certainly intend that (1) and (2) do. But a *maxim* that predicates self-interest of a permanent rule of promise-keeping is materially inconsistent, and a *maxim* that both asserts and denies the same state of affairs is logically inconsistent (other examples of the latter might be, "Out of an impulse to make mischief, I decline to state any of my maxims, in order to disturb the repose of my readers;" "From iconoclasm, I make it a rule to ignore rules, in order to supply the authorities with something to do;" and so forth). We can conceive without contradiction communities in which everyone acts on these maxims, yet we cannot conceive without contradiction the content of these maxims themselves.

However, I make no claim that yet other, morally controversial kinds of maxim also must necessarily fail to satisfy some of the consistency constraints that define a genuine preference, because I do not concur with Kant's belief that a single, well-defined normative moral theory can be read off from rationality criteria more generally; more on this in Section 5 below. My purpose here is merely to suggest some ways in which reasons for actions can also be causes of action without being or including desires. An intellectually committed Egoist who must systematically override his own altruistic or generous desires in order to honor the maxims that express his considered moral convictions would provide another example of an agent whose actions are motivated by reasons that are not desires.

Baron speaks as though there is a conflict between "act[ing] on maxims" and acting "from motives;" but there is no such conflict. Maxims themselves become motives when we are motivated by conceptually recognizing their rational intelligibility to act on them. An intention whose intentional content satisfies the consistency constraints that establish its rational intelligibility within the concepts constitutive of our perspective, and therefore its status as a genuine preference, provides both the intentional content capable of guiding our action, and also our motivationally effective reason to so act. A motivationally effective maxim is then a genuine preference that both describes and also motivates the action that conforms to it.

Now we have seen in Section 3 that Baron's claim that Kant does not need primary motives in order to explain how duty guides action is backed by an analysis of secondary motives that is supposed to do this work. But we have also seen that that analysis cannot stand alone independently of any reliance on primary motives, for in that case there is no way of motivating the individual act-tokens that, over time, both habituate one to the long-term course of conduct that secondary motives identify, and also constitute that long-term course of conduct itself. Without primary motives to undertake the project of moral self-improvement – to practice acting in accordance with duty, to occurrently reflect on what respect for the moral law requires of us on a particular occasion, to cultivate moment-to-moment sensitivity to those requirements, and so forth – it is hard to see how duty as a secondary motive could ever take hold. Once it does take hold, it will bear a relation of instantiation to individual act-tokens most of the time, even if the requirements of duty are violated occasionally, and even if one then need not be preoccupied with these requirements most of the time; nothing in Baron's account of duty as a primary motive requires that we never fail to act on this motive, nor that we constantly obsess about what duty

requires on each occasion of action. To achieve this degree of habituation is not to dispense with duty as a primary motive, and I have suggested that it is implausible to think one could. A viable Kantian theory of agency needs primary motives as much as any other kind, because it is not exempt from the requirement of identifying the sufficient condition that causally precipitates the intentional physical behavior of an agent who acts for reasons. I attempt to meet this requirement in the following sections.

4.3. Minimally Precipitating Thoughts

An occurrent thought or belief qua psychological and neural event *minimally precipitates* action if it does so without directing and guiding it, as when the thought of the ungraded stack of exams on my desk causes me to engage in elaborate rituals of filing, pencil-sharpening or house-cleaning. The content of this thought to some extent determines my actions in the minimal sense that different content would have precipitated different actions. But it does not direct and guide those actions. It is not reflected in the content of the actions I perform.

It is tempting to explain such behavior by stipulating an intermediary occurrent desire to avoid grading the exams. Such a desire may or may not be present. But even where it is, it is an open question which event – the occurrent thought-event or the occurrent desire-event – is doing the causal work. I offer some reasons for this in Sections 4.4.2-3, below. Hence the explanation of such actions as simple "avoidance behavior," which presupposes the existence of a motivationally overriding aversion or aversive desire, may be factually inaccurate.

However, the causal connection between what I think or believe and what I do may be even looser and more free-associational than this: any sufficiently vivid, occurrent thought may suffice to cause me to roll out of bed and begin my day. These are both cases in which the content of my thought or belief bears little or no conceptual relation to the action it causes. The thought- or belief-event is causally contiguous to the action-event; over time, a constant conjunction between them may even be observed, and a causal connection justifiably inferred. Lacking, however, is any governing role for the content of my occurrent thought or belief as directing and guiding my action. This lack characterizes those occurrent thought- or belief-events that are minimal precipitators of action.

4.4. Will

My interest in the motivational efficacy of reason on action restricts my attention in what follows to occurrent thought- or belief-events whose conative power is heightened by content that does govern, direct and guide action – the intentional content of which reflects this. I focus in particular on the content of occurrent abstract thought- or belief-events formed by the intellect. By the *intellect* I mean the mental capacity for abstract thinking: for deductive and inductive reasoning and analysis, formulating generalizations and universal rules and principles, and so for hypothesis construction and theory-building. Sometimes I use the term *intellection* as a shorthand referent to these capacities. The intellect, then, is the capacity for theoretical

rationality. The ability to perform action governed, directed and guided by the rational content of an occurrent abstract thought or belief of the intellect is what Kant calls "the capacity to act in accordance with [one's] representation of laws – that is, in accordance with principles …" [G, Ak. 412] Following Kant, I define *will* as intellect that is causally effective in this sense [2C, Ak. 89]. If will is causally effective intellect and if intellect is the capacity for theoretical reason, then theoretical reason can be causally effective in precipitating action. The following analysis therefore rejects any reified distinction between theoretical and practical reason. Sometimes reason – i.e. the intellect – is causally effective, sometimes it is not. Reifying intellect into "practical reason" when it is causally effective and into "theoretical reason" when it is not violates Occam's Razor, and leads to misunderstandings. For it is the causal efficacy of only one capacity, not two, that is at issue:

> Solely if pure reason of itself can be practical and really is, as the consciousness of the moral law proves it to be, is it always just one and the same reason which, whether for theoretical or for practical intent, judges a priori in accordance with principles. [2C, Ak. 121]

4.4.1. Motivational Ineffective Intellect

An agent's intellect can be causally effective to varying degrees, and so the will can possess varying degrees of strength or weakness. The intellect is *motivationally ineffective* when the agent occurrently thinks or believes something of an abstract nature that under ideal circumstances could but under actual ones does not move the agent to act. For example, consider further the case discussed in Volume I, Chapter VII.3.4, of dissociation from Nagel's impersonal standpoint. I might entertain the abstract belief that

(3) Someone should clean up this neighborhood.

But I might nevertheless fail to be galvanized into action by the recognition that

(4) That someone is me.

In such a case we commonly say that the agent suffers from weakness of will. Weakness of will can be of two kinds. The intellect may be simply impotent, a weak and shallow whisper inadequate to catch the agent's attention or overcome her inertia. Alternately, the intellect may be not impotent as such, but instead merely subverted by stronger countervailing desires that drown it out. In this case the agent's will may be simply upended by desire, or intellectually reconfigured so as to justify desire, producing pseudorational apologia and ideologies that merely rationalize desires that would be motivationally overriding even without them. In this case the agent manipulates her intellect, and its rational capacities, in the service not only of pursuing desire-satisfaction, but of justifying its pursuit to conscience and to others. I discuss this latter case at greater length in Chapter VII.

4.4.2. Opportunistically Effective Intellect

Second, an agent's intellect may be *opportunistically effective* to varying degrees, when he occurrently believes something of an abstract nature that is again capable of moving him to act, but actually does so only to the extent that other circumstantial factors – desires, emotions, habits, external conditions – do the heavy lifting; this is what Baron calls a "hybrid motive" (7; also see the excellent discussion at 151-156). Kant's specifically moral example is of the grocer who believes that prices should be fairly set, and does, indeed, charge his customers fairly – because this enables him to retain their business [G, Ak. 397]. In this case self-interest provides the impetus to perform an action the agent may believe is worth performing anyway, but otherwise would lack sufficient personal incentive to perform. But Kant's distinction between acting from duty and acting merely in accordance with duty is a special case of a more general distinction that has broader, nonmoral application as well: between acting from reason and acting merely in accordance with reason. To *act from reason* would be to perform an act-token whose intentional content instantiates the rational content of the occurrent abstract thought or belief that precipitated it:

(5) ∴ I will clean up this neighborhood.

(5) expresses the intention behind an action that is governed, directed, and guided by reason. By contrast, an opportunistically effective intellect performs action whose intentional content does, indeed, instantiate the rational content of the agent's occurrent abstract thought or belief; but that was precipitated by an incentive other than the content of that occurrent abstract thought or belief. In that case, one acts merely *in accordance with reason*: I clean up the neighborhood as (3) requires, but only in order to avoid grading the stack of exams on my desk. Here I heed one set of rational principles, but only in order to avoid heeding a different one.

Consider a second example. Lucille's occurrent abstract beliefs that

(6) a rational agent obtains all essential nutrients;

(7) certain essential nutrients are found only in broccoli;

(8) ∴ a rational agent includes broccoli in her diet

may have enough conative force to draw her attention, but by themselves not enough to move her to instantiate (8) in her own actions. However, the following conditions may provide the needed additional impetus: first, Lucille is hungry; and second, there is nothing else to eat. Having added broccoli to her diet under these unusual conditions, Lucille then may be able to cultivate the habit of eating broccoli on a measured, daily basis even in their absence. Even though her behavior then conforms to what reason requires, reason itself was parasitic on other incentives, motivating her only to use to her rational advantage the internal craving and external deprivation that precipitated her response. In this kind of case, the intellect uses contingent conditions opportunistically, in order to fuel action to which it ascribes independent value.

4.4.3. Motivationally Effective Intellect

Finally, an agent in fact may act from reason: an agent's intellect is *motivationally effective* when the rational content of its occurrent abstract thoughts or beliefs moves the agent to perform action whose own intentional content instantiates the rational content of the occurrent abstract thought or belief that precipitated it, regardless of other desires or inclinations, and whether other internal or external circumstantial factors cooperate or not. I opened Chapter I with the question whether moral principles could have any further effect on human behavior beyond the pseudorational one described in Section 4.4.1 above, and answered that question in the affirmative in Chapter I.7.3.3. To defend that answer successfully requires making the case for a motivationally effective intellect. The rest of this chapter addresses this task in depth.

Ordinarily we describe this kind of case as one of strength of will, meaning that intellectual convictions prevail over recalcitrant desires, emotions, or external conditions in causing one to act. However, this conception of strength of will conceals a further distinction between three kinds of case.

(1) The first is the case in which motivationally sufficient thought or beliefs are present, motivationally sufficient motives of other kinds – desires, self-interest, impulses, drives – are also present, the two conflict, and the thoughts or beliefs override the other, conflicting motives.

(2) Then there is the case in which motivationally sufficient thought or beliefs are present, motivationally sufficient motives of other kinds – desires, self-interest, impulses, drives – are also present, and the two kinds of motives are not conflicting but rather complementary. In this case either kind of motive would have been sufficient to cause the action in the absence of the other, but as it happens, both kinds of motive are operative; this is what Baron calls an "overdetermined action." In the end, Baron denies that the notion of an overdetermined action is even intelligible on Kant's view. (159-162) She says,

> Suppose I believe that something is required of me, and act accordingly. If I act from the thought that it is required, it does not make sense to say that I may at the same time be acting from other motives. (159)

But I do not see why not. By analogy, suppose only two aspirins were required to cure your headache but you take four simultaneously, just to make sure. If two of the four were wasted, which two would that be? Or would it be more accurate to say that 50% of each of the four was causally effective in curing your headache, while the remaining 50% of each was wasted? Neither seems as accurate as saying simply that you cured a headache with four aspirins that could have been cured with two. Does it matter whether the multiple causes all involve intentional states, as in the passage above? Baron argues that it does, because both causes must be reasons:

> We can imagine that [a committed philosophy major] takes [the course] because he wants to, and that it is also true that if he did not want to take it, he would take it anyway because it is required. But that does not make his action overdetermined. He is taking it because he wants to …; in other circumstances he would take it because it is required. But he does not take it for *both reasons at once.* (161)

Baron's argument here is that a belief and a desire cannot conjointly overdetermine an action because both must provide reasons for action, and both reasons cannot be true of the agent at once. But again I do not see why not. I do not see why I cannot take a course both because it is required and also because, in any case, I want to; or, alternately, both because I want to and also because, in any case, it is required. I do not think all occurrent causes of action are reasons, or even that all simultaneous, intentional motive causes of actions must be reasons. But I see no reason why they might not be in a particular instance.

(3) A third case is that in which the choice is not between two conflicting actions motivated by two conflicting conative sources, but rather between action motivated by reason, and inaction. In this case desire, emotion, or external circumstance, whether conflicting or not, is of negligible impact, assuming it is present at all. If the intellect does not motivate a particular action, the agent does nothing; and whatever action the agent does perform was motivated by intellection. I examine this third case at length in order to see more clearly what it is, precisely, that overrides inclination in the first and operates concurrently with inclination in the second.

4.5. Fully Effective Intellect and Implicit Self-Recognition

This third case is the one Kant describes in the *Groundwork* as a "perfectly good will," i.e. one that is "infallibly determined" by reason, such that "the will is then a capacity to choose *only that* which reason independently of inclination recognizes to be practically necessary …." [G, Ak. 412], and is "necessarily submissive to rational considerations [*Gründe der Vernunft*] in accordance with its nature." [G, Ak. 413] Kant intended this concept to address specifically moral actions motivated by reason. In this section I mean to address any action motivated by reason, and mark this intention by speaking not of a perfectly good will, which implies moral evaluation, but rather of what I shall call a *fully effective intellect*. How might we understand such an intellect that, in virtue of "its own nature," only and always chooses to act according to what it recognizes as rationally required "independently of inclination"?

I say more of an exegetical nature about this passage in Section 5.1, below; and even more elsewhere.[12] For present purposes note merely that the foregoing account of recognition developed in Chapter II, above, explains what it might mean to say that in a fully effective intellect "reason … recognizes" an action as "practically necessary." There I appropriated Kant's technical analysis of recognition as the ability to identify something at any given moment as the same, with respect to some property, as that which was cognized earlier; that is, to subsume it under a concept. Then reason recognizes an action as practically necessary if it subsumes the action under the concept of what the agent is practically required to do.

An agent with a fully effective intellect necessarily, by definition, instantiates in his own action the descriptive principles of rational action he believes to be true. 4.4.2.(6) –(8) above would be examples of principles descriptive of rational agency that an agent might believe to be true. An agent who occurrently believes these principles and has a fully effective intellect instantiates in his behavior these rational principles *because* he occurrently believes them to be

true. That is, he not only recognizes the principles as true; he implicitly recognizes them as true *of him*. He implicitly recognizes himself as a specific instance of a more general concept within his perspective – the concept of a rational agent – that in turn figures in a descriptive principle he occurrently believes, of what a rational agent does under his particular set of circumstances. The descriptive principle he occurrently believes defines for him the rational action his current circumstances require and which he therefore performs. "Rational" here is shorthand for the account of rational action summarized in Sections 1 – 4, and elaborated in Chapters II and III, above. An agent with a fully effective intellect is motivated, by the occurrent thought of the principles of rational agency in which he implicitly recognizes himself and whose truth he believes, to instantiate those principles in his actions. Next I sharpen this thesis by contrast with certain other ones with which it might be confused.

In Volume I, Chapter VI.1, I defined a motive as *self-interested* if it includes an interest the self takes in its own condition. Does the foregoing account of implicit self-recognition qualify as a self-interested motive according to this definition? No, because an agent who implicitly recognizes herself in a descriptive principle of rational agency does not thereby take an interest in her own condition – anymore than we take an interest in our own condition as readers by virtue of implicitly recognizing ourselves in a general description of readers. Implicit self-recognition does not imply self-evaluation or self-concern, whereas self-interest does. A fully effective intellect that always acts from and in concert with rational principle embodies rationality in her action. So the need for evaluation, and for the independent standard of comparison that evaluation requires, do not arise. This connection between principle and action is "practically necessary," to use Kant's words, because it defines not only what a rational agent is conceptually, but thereby what such an agent does in practice:

> Everything remaining that is psychologically dependent on this concept, i.e. in so far as we empirically observe these capacities of ours in their exercise, is abstracted out (for example, that human understanding is discursive, [that] its representations are thoughts and not intuitions, that these follow one another in time, that its will is burdened by being dependent for its contentment on the existence of its object, etc., none of which can be true in the highest being); and thus nothing more remains from these concepts through which we think a pure rational being than just what is required by the possibility of thinking a moral law. [2C, Ak. 137]

That she always acts on rational principles means that she acts in character, i.e. she is "necessarily submissive to rational considerations in accordance with [her] nature." [G, Ak. 413].

Nor does this account of implicit self-recognition imply that a fully effective intellect acts *in order to confirm the truth of* the rationality principle that governs, directs and guides its action. The relation between the agent's occurrent thought of the principle and the action that expresses it is neither instrumental nor verificatory. It is closer and more unmediated than that. Recall from Section 2, above, that the highest-order disposition to literal self-preservation is a motivationally overriding disposition to preserve the theoretically rational coherence of the self.

In practice this is a disposition to *do* whatever it takes to preserve the horizontal and vertical consistency over time of the experiences one has. Actions one takes are among the experiences one has. So a fully effective intellect has a highest-order, motivationally overriding disposition to act in ways that preserve internal consistency; in effect, a motivationally overriding disposition to heed the conclusions of theoretical reasoning it reaches, and to instantiate them in its own behavior. A fully effective intellect embodies and enacts in practice the descriptive principles of rational agency it occurrently believes to be true. I shall say that it has a *rule-governed disposition to rational action* that is actualized when prompted by its occurrent thoughts and beliefs. Thus it can be viewed as a generalization of McClennen's conception of a rule-guided disposition to resolute choice.

Finally, implicit recognition of oneself in a principle of rational agency is different from contingent identification of oneself as, for example, a baby-boomer or jazz aficionado, in which one might psychologically invest or divest. One can continue to be oneself, and to be rationally intelligible to oneself, even after having divested oneself of the identity of jazz aficionado. By contrast, an agent with a fully effective intellect could not continue to be himself, or to be rationally intelligible to himself, after having ceased to implicitly recognize himself in the principles of rational agency on which he acts. We saw in Chapter II.3 that to make something rationally intelligible to oneself is to recognize it as an instance of some higher-order concept. To make oneself rationally intelligible to oneself, then, is to implicitly recognize oneself as an instance of some higher-order concept.[13] Many such higher-order concepts are available. But the concept of oneself as a rational agent is one that an agent with a fully effective intellect finds impossible to avoid, because all of his actions instantiate it. So such an agent implicitly recognizes himself in that one, in virtue of making his actions rationally intelligible to himself in the first place. To continue to act rationally in the sense explained without implicitly recognizing himself as rational would be to violate the requirement of vertical consistency, and so would engender the self-contradiction that this account of rationality excludes. Were a rational agent unable to implicitly recognize himself as rational, he would be unable to make himself – or, therefore, the rest of his experience, including his actions – rationally intelligible to himself at all. Failure of implicit self-recognition would be equivalent to failure of rationality and therefore failure of agency for a fully effective intellect.

To the extent that we make things rationally intelligible to ourselves, then, and making things rationally intelligible implies rational agency, as Chapter II.6 argued, we are rational agents. I have argued here that for a fully effective intellect, rational agency implies maximal attention (or *Achtung*, as Kant would put it) to the conclusions of theoretical reasoning; and so their unobstructed motivational influence on action. A fully effective intellect that implicitly recognizes itself in descriptive principles of rational agency precipitates action governed, directed and guided by those principles. It thereby actualizes and instantiates its rule-governed disposition to rational action.

Thus implicit self-recognition in occurrently thought or believed principles of rational agency is a direct, practical application of the requirement of vertical consistency to the relation between intellection and action. Intellection supplies the descriptive principles, action supplies their instance; intellection supplies the cause, action the effect; intellection supplies the cue, action the actualized disposition. This is the sense in which an agent with a fully effective intellect acts "in accordance with [her] representation of laws – that is, in accordance with principles …." [G, Ak. 412]. The rational content of the principle she occurrently thinks or believes activates her rule-governed disposition to enact this principle in her action. She thereby reinforces her own highest-order disposition to literal self-preservation, and so the rational coherence of her self.

5. An Instantiation: Kant's Descriptive Moral Theory

Next I apply this analysis of ideally rational motivation in general to the case of ideally moral motivation in particular, using Kant's normative moral theory as an example. My remarks are intended to have application to any normative moral theory, not only Kant's. But since Humean moral theories are all Instrumentalist in structure, and since all Instrumentalist principles are conditionals that embed categorical principles in their consequents, it will be convenient to dissect a theory that consists exclusively in such categorical principles. Kant's is the most sophisticated theory of this type we have available.[14]

5.1. Descriptive

In this section I extend Kant's own characterization of what it means to be motivated by reason alone to perform specifically moral action, through an analogy with the de facto motivational efficacy of certain non-moral principles of reasoning. In Section 5.2, following, I show that Kant's descriptive moral theory is genuinely explanatory, and so qualifies as theoretically rational in the sense already explained.

Kant himself goes further: he believes that normative moral directives can be logically *derived* from principles of theoretical reasoning, so he often equates "rational" and "moral." As indicated in the preceding chapter, I do not share this belief. The success with which the obligation of promise-keeping was derived from McClennen's concept of resolute choice depended on relinquishing rationalist aspirations, whereas Kant's derivations depend on them. I believe that some substantive moral principles may instantiate universal rational ones without being logically implied by them, because no universal principle logically implies all of its instances. However, I sometimes use Kant's equation of "rational" and "moral" in the present section, in explicating Kant's own view.

Kant's moral theory is often regarded as inherently prescriptive. Elsewhere I defend the view that this can be explained by a misinterpretation of Kant's concept of *Achtung*. Here I merely reconsider at greater length the passage quoted above from the *Groundwork of the Metaphysic of Morals* in which Kant gives the notion of a perfectly rational being a moral inflection, in order to explicate the status he takes a moral theory to have. He says,

> If reason unavoidably determines the will, then in a [perfectly rational] being of this kind the actions which are recognized to be objectively necessary are also subjectively necessary, that is, the will is then a capacity to choose *only that* which reason independently of inclination recognizes to be practically necessary, that is, to be good. [G, Ak. 412; italics in text]

As Kant sees it, a perfectly rational being is motivated, by the occurrent thought of the actions her moral principles rationally imply, to perform those actions. Thus like a fully effective intellect more generally, Kant's perfectly rational being is causally determined by reason to act in this way: reason has the same motivational function for a perfectly rational being that any overriding disposition to behave has for us. An analogy with one particular disposition may be useful: In Kant's perfectly rational being, reason determines action just as necessarily as in us, reason determines our inference from P and if P then Q, to Q.

Two features of this analogy are important for understanding the motivational psychology of Kant's perfectly rational being. First, acting on the promptings of reason is a natural and unforced expression of her rule-governed disposition to do so. Because her action is unconstrained by doubt, inhibition or conflict, it expresses a certain kind of freedom. This is what it is like for Kant's perfectly rational being to act – either mentally or physically – in accordance with her conception of what reason requires: It is a natural and unforced expression of her rationality, free and unconstrained by doubt, temptation, or narrow personal agendas. This is the state that Kant describes as one of "subjective necessity."

Second, inferring Q from P and if P then Q is something we just cannot help doing. Of course we can deliberately declaim *modus ponens* incorrectly, or make a genuinely mistaken inference if the argument is extended and the premises buried deeply in turgid text. But once we see the premises and the structure of the argument, it is not seriously open to us – nor does it normally occur to us – to infer not-Q from P and if P then Q. Our implicit recognition of the objective validity of *modus ponens*, and of ourselves as reasoning validly, determines our adherence to it. This is what Kant means by "objective necessity."

Subjective and objective necessity coincide when what we are naturally and freely disposed to do coincides with what we are and recognize ourselves as being objectively required to do. They fail to coincide when what we regard ourselves as being objectively required to do interferes with or inhibits our natural dispositions to act – as reason sometimes does our empirical inclinations. In the behavior of Kant's perfectly rational being, objective and subjective necessity coincide because the very idea of the rational action itself has an overriding motivational force that it does not occur to a perfectly rational being to try to resist. Kant's perfectly rational being, then, both freely expresses her rationality in action, and is fully determined by it.

For Kant, rational principle has objective necessity for two reasons. First, rational principle is objective in the sense that we conceive it as having universal applicability. This is just to say that we conceive it as a law of logic.[15] Second, it is objective in the sense that it presents to

us the same imperturbable, unquestionable inexorability as does any event or state of affairs in the world whose internal logic is independent of our wishes and our existence: If P and if P then Q, then Q, whether or not we believe it and whether or not we exist; and similarly, for Kant, with the deliverances of moral principle. Because Kant views moral principles as implied by principles of logical reasoning, he ascribes to moral principles the same epistemic status and psychological impact as a more general principle of reasoning such as *modus ponens* ordinarily has. Kant's perfectly rational being is one for whom the objective necessity of moral principle is itself a subjective expression of freedom. In this sense, moral principle describes the unconstrained behavior of Kant's perfectly rational being.

This means that prescriptive "ought"-language is not inherently a part of moral theory for Kant. For a perfectly rational being, moral theory supplies a straightforwardly descriptive account of her behavior, relative to which "ought"-language is otiose. Kant states this conclusion explicitly when he says, in Chapter III of the *Groundwork*, that

> this 'I ought' is actually an 'I will' that is valid for every rational being, under the condition that reason in him were practical without any hindrance. For beings who, like us, are also affected by sensibility as motives of a different kind – beings for whom what reason by itself alone would do does not always occur, this necessity of action is called an "ought," and the subjective necessity is distinguished from the objective. [G, Ak. 449]

Although this idea of moral theory as ideal descriptive theory is elaborated at greatest length by Kant, it is not unique to Kant's thinking. It is implicit in Aristotle's suggestions that just and temperate action is defined by the standard set by individuals who actually are just and temperate,[16] and that the excellent individual sets the standard for judging what is truly good and pleasant.[17] And it is given a somewhat more extended treatment in Sidgwick's analysis of "ought". Sidgwick distinguishes two uses of "ought". The first is the narrow sense in which we judge what an individual ought to and therefore can bring about through his own volition. But there is a wider sense in which "the word merely implies an ideal or pattern which I 'ought' – in the stricter [i.e. the first] sense – to seek to imitate as far as possible,"[18] even though I do not, in using the word in this wider sense, imply any ability to bring about the ideal through my own volition. We use the word in this wider sense when we talk about what ought to be the case, or what I ought to know or feel. "In either case," Sidgwick tells us, "I imply that what ought to be is a possible object of knowledge: i.e. that what I judge ought to be must, unless I am in error, be similarly judged by all rational beings who judge truly of the matter."[19] That is, the ideal to which I bear the "ought" relation is one that describes a possible, objectively knowable state of affairs. Sidgwick then goes on to give a Kantian analysis of the narrower sense of "ought" as having motivational force exclusively for agents who experience conflicts between impulse and reason, and says no more about the wider sense. Nevertheless, his notion of the use of "ought" as implying a relation between an imperfect human being and an ideally described state of affairs captures the kernel of Kant's conception of moral theory as an ideal descriptive theory. Hence this conception of ideal descriptive moral theory is one even a Humean can embrace. I defer to

Chapter IX an explanation of in what the compulsionary "ought" relation to any such theory consists.

5.2. Explanatory

What does an ideal descriptive moral theory look like? And is it really a theory, rather than, say, a set of loosely consistent beliefs or presuppositions of action? In this section I argue that Kant's descriptive moral theory satisfies many of the criteria of a genuinely explanatory theory, and so instantiates the more general rationality principles I defended in Chapters II and III. There is no inconsistency in contending that Kant's ideal descriptive moral theory is both explanatory and normative for human beings. To repeat a point made in Volume I, Chapter III, 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. 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. Kant's ideal descriptive moral theory has exactly this form. Hence I take on Rawls' ambition, which drove his early work in metaethics,[20] to provide moral philosophy with a thoroughgoing analogy with science. However, because I take issue with Kant's much stronger claim that his moral theory is implied by rationality principles, my analysis of a fully effective intellect requires only that Kant's moral theory instantiate those principles; and so that a similar analysis be available to any developed normative moral theory that claims practical application. This in effect stipulates rationality criteria that any viable normative moral theory must meet.

Following are some principles from what I call Theory *K*, in honor of its Kantian origins:

(A) (1) If a rational being has the opportunity and desire to commit suicide, she refrains from it. [G, Ak. 422, *passim*][21]
(2) If a rational being makes a promise, he keeps it. [G, Ak. 422, *passim*]
(3) If a rational being has natural talents, she sometimes cultivates some of them. [G, Ak. 423, *passim*]
(4) When a rational being encounters individuals in need, he sometimes helps some of them. [G, Ak. 423, *passim*]

(B) (1) When a rational being is moved to act, she performs only those acts that can be willed as a universal law of nature. [G, Ak. 402, *passim*]
(2) When a rational being is moved to act, he performs only those actions consistent with treating humanity as an end in itself. [G, Ak. 427, *passim*]

(C) (1) When a rational being acts, she is motivated by respect for the moral law. [G, Ak. 400, *passim*]
(2) When a rational being resolves to act, his will makes universal law. [G, Ak. 431, *passim*]
(3) When a rational being resolves to act, she legislates autonomously for a kingdom of ends. [G, Ak. 433, *passim*]
(4) When rational being acts, he is noumenally free and phenomenally determined. [G, Ak. 451, *passim*][22]

(D) (1) The causality of the will of a rational being is expressed in action performed out of respect for the moral law. [G, Ak. 453, *passim*].
(2) The freedom of the rational being as noumenal subject is expressed in such moral action. [G, Ak. 454, *passim*]

Principles (A.1) through (D.2) make purely descriptive claims about the behavior of certain sorts of phenomena, namely rational beings. The concept of a rational being can be rendered in similarly descriptive terms, i.e. as a being possessed of theoretical and practical rationality. Hence (A.1)-(D.2) contain no prescriptive terms. Can they be said to form part of a genuinely explanatory theory? I think they can, if we loosely apply the conventional criteria that identify a set of principles as a theory. Whether *K* is a good or the correct theory is, of course, a separate issue. I defer discussion of *K's* explanatory adequacy to Chapter IX.

First, a theory begins with *hypotheses*, i.e. proposed lawlike explanations of phenomena that are accepted conditionally on confirmation of their experimental predictions. From such a hypothesis, we should be able to infer causal regularities that can be experimentally tested. The more confirmable predictions we can make, the more credibility accrues to the hypothesis. Take (A.2), above. (A.2) (together with the suppressed premise that one's parents are rational beings) implies that if one as a child asks one's parents to attend school events in which one plays one's tuba – i.e. school recitals, dances and the marching band, and they agree to do so, then, barring unforeseeable catastrophes, they will attend the school recital to hear one play one's tuba. If, under these circumstances, they actually do attend the school recital, then one has confirmed at least one experimental prediction of (A.2). A second such prediction might be that, if, as a teenager, one confides in one's friend Millicent about one's crush on Conrad, then Millicent will refrain from apprising all one's friends of one's feelings about Conrad. If Millicent does, under these circumstances, keep mum about one's crush on Conrad, then one has some confirmation of a second experimental prediction of (A.2). A third might be that, as an adult, if Angus accepts one's job offer to join one's firm as a vice-president, Angus will in fact discharge his vice-presidential responsibilities. Angus does in fact discharge his vice-presidential responsibilities, and lo! One has confirmed a third experimental prediction of (A.2). Notice that all the principles in group (A) are susceptible to the same sort of experimental testing.

The more confirmation accrues to (A.2), the more one is entitled to regard (A.2) not just as a hypothesis, but as a *law*, i.e. a true hypothesis stated in the form of a generalization that links states of affairs causally, as does (A.2): When a rational being makes a promise, she will keep it. Like all the principles in group (A), (A.2) satisfies the *nomological* requirement that it support counterfactual conditionals: If a rational being were to make a promise, she would keep it, and if she had made one, she would have. Like McClennen's pragmatist model of resolute choice, (A.2) ranges over not only the actual past, present, and future, but over possible pasts, presents, and futures as well; it has universal rather than merely spatiotemporally limited application. Thus (A.2) stands in contrast to a mere *accidental generalization* like

(E) If someone keeps his promises, he is a rational being,

since someone could conceivably keep his promises – say, because he had been hypnotized into doing so, without being a rational being.

It might be objected that suppressed premises of the sort mentioned above, that the individual in question is a rational being, are themselves accidental generalizations over instances of behavior that happen to, but may not in all foreseeable cases evince rationality. But this objection could be raised as well of any suppressed premise that identifies an event or state of affairs in its subject term: "This object is a paraffin candle" is similarly interpretable as an accidental generalization over instances of object behavior that may not in all foreseeable cases evince paraffin candlehood (perhaps it will bob about when thrown into boiling water, like plastic, instead of melting). Hence these two kinds of suppressed premise must stand or fall together.

Similarly, it will not do to object that, unlike scientific laws, (A.2), and indeed all the propositions in (A), are true by stipulative definition of "rational being", since the same objection could be raised about the status of the hypothesis, "Paraffin melts when put into boiling water:" If it doesn't melt, then either we were wrong about what that substance is, or we were wrong about how paraffin behaves. Again both kinds of hypothesis must stand or fall together.

Straightforward theories contain both lower-level and *higher-level* laws. The latter are laws that satisfy the same criteria just discussed, but that generalize over lower-level laws with respect to more abstract features of the phenomena described. According to Kant's reasoning, the two propositions collected under (B), above, are of this kind. (B.1) and (B.2) are themselves laws from which (A.1-4) can be deduced as experimental predictions: For example, if a rational being who is moved to act performs only those acts that treat humanity as an end in itself, then if such a being makes a promise, she will keep it (because keeping one's promises is an act that treats humanity as an end in itself). Let us try to fit (A) and (B) into a *Hempelian covering law schema* of the following sort:

(Covering Laws) $L_1, L_2, \ldots, L_n$

} Explanans

(Particular Circumstances) $C_1, C_2, \ldots, C_m$

_____

(Phenomenon to be explained) E } Explanandum

Although we may disagree with the details of Kant's own reasoning about the practical implications of his various formulations of the Categorical Imperative, the Hempelian schema organizes some of the propositions of Theory *K* rather well:

(F) (L_1, L_2) If a rational being makes a promise, he will keep it (A.2); and when a rational being encounters individuals in need, he will sometimes help some of those individuals(A.4);[23]

(C_1, C_2) You asked your parents if they would attend school events to hear you play your tuba, and they said they would;

_____

(E_1) Your parents attend the school recital to hear you play your tuba.

(G) (L_3, L_4) When a rational being is moved to act, she performs only those acts that can be willed as a universal law of nature (B.1); and that treat humanity as an end in itself (B.2);

(C_3, C_4) Keeping one's promises and sometimes helping some of the needy can be willed as universal laws of nature (B.1), and also treat humanity as an end in itself (B.2);

_____

(E_2 (= (L_1, L_2)) If a rational being makes a promise, he will keep it (A.2); and when a rational being encounters individuals in need, he will sometimes help some of those individuals (A.4).

In the *Critique of Pure Reason*, Kant refers to a chain of deductive inferences like the sequence {(1), (2),..., n} as the "ascending [as opposed to the descending] series of syllogisms of reason" [1C, A 331/B 388], i.e. that series the members of which increase in generality and comprehensiveness relative to the particular facts (or "empirical conditions") with which it begins. About the ascending series Kant also claims that in order to generalize over such laws or premises to increasing degrees of abstraction, we have to assume a totality of such laws, and their termination in what Kant calls transcendental ideas of reason. [1C, A 336/B393 – A 337/B 395; also cf. the section on "The Regulative Employment of the Ideas of Pure Reason"]. Indeed, Kant thinks that we must assume such a unified theory in order to exercise our reason and

understanding in the search for empirical truth at all [1C, A 651/B 679]. This is the core idea of the requirement of vertical consistency developed in Chapter II.

In the language of scientific theory, this would be to assume the internal coherence and completeness of the theory, and the termination of its higher-level laws in the *theoretical constructs*[24] of the theory and the principles governing it. The theory, in this parlance, is a conceptually higher-level hypothesis that is accepted as true because it successfully explains lower-level, law-governed uniformities as manifestations of "deeper" and (according to some) unobservable entities and processes that are themselves governed by theoretical laws and principles. Examples of such constructs from Theory *K* appear with increasing frequency as the level of abstraction of the principles increases: "Reason", "will", "law", "humanity", and "end" are theoretical constructs in (B), according to this description, as are "respect," "kingdom of ends," "freedom", and "noumena" in (C). All are abstractions that combine to form an ideal type[25] whose behavior explains the uniformities of behavior of perfectly rational beings as described in (A).

These theoretical constructs are, like scientific theoretical constructs, governed by two kinds of principles. First, there are *internal principles* that describe their behavior: The categorical imperative describes the operation of the rational will as legislating the moral law; Kant's account of the activity of reason in the Dialectic of the first *Critique* explains how the ideas of humanity as an end in itself and of the kingdom of ends function for us, and, together with the *Groundwork*, in what freedom, autonomy, and the noumena-phenomena distinction consist. I elaborate these principles at length elsewhere. My aim here is to use Kant's moral theory to illustrate how and where internal principles that describe the behavior of such theoretical constructs are to be found in a moral theory.

In addition to internal principles, *K* also contains *bridge principles* that connect these constructs with familiar empirical phenomena. (D.1) and (D.2) are bridge principles. Both contain what we might describe as "double connections": First, there is the causal double connection in (D.1), between (a) the causality of the will and the feeling of respect, and (b) the feeling of respect and the resulting moral action: Rational principles of action elicit respect, which in turn motivates moral action. Second, there is the evidential double connection in (D.2), between (a) freedom and the noumenal subject, and (b) the noumenal subject and the moral action: Freedom is manifested by a subject whose behavior is not determined by empirical inclinations – i.e. a noumenal subject, and noumenal subjecthood is evinced by moral action. In both cases, these principles link the moral actions we observe with the theoretical constructs that ultimately explain them. So far, at least, Theory *K* does seem to satisfy at least some of the rudimentary requirements of a genuine theory.

The concept of a rational being that Kant deploys in Theory *K* is not shorthand for the exact same analysis of rationality that I have offered in Chapters II and III. Kant's concept presupposes something like that analysis, but claims to logically imply many more specific descriptive principles about the moral actions a rational being performs. This is where I part company with Kant. If no universal principle logically implies all of its instances, then it is

unlikely that a sufficiently comprehensive analysis of rationality can logically imply particular moral principles that all fully rational agents necessarily follow in virtue of their rationality. However, such particular moral principles might well instantiate more general rational ones. Although Kant's ideal descriptive moral theory has universal application to all rational agents, not all rational agents need occurrently believe the descriptive principles of his theory to be true, nor implicitly recognize themselves in these descriptions. Hence Theory *K* may not have the same causal efficacy for all motivationally effective intellects – or even for all fully effective intellects.

However, so long as there is some ideal descriptive moral theory whose principles a rational agent occurrently believes, and in which such an agent implicitly recognizes herself, such that this theory instantiates and satisfies the more general rationality requirements outlined above, such a theory qualifies as rational. And then there will be some such theory that has causal efficacy for any motivationally effective intellect; i.e. some such actions "which reason independently of inclination recognizes to be practically necessary, that is, to be good."[G, Ak. 412] Nevertheless, I enumerate some further, non-ideal criteria a rational moral theory must satisfy in Chapter X below, and argue that only Kantian-*type* moral theories satisfy them.

6. Two Ideals of Rational Motivation

So far I – and Kant – have characterized such inclinations, whether conflicting or harmonious, as of negligible impact on action. A fully effective intellect, a perfectly good will, a fully rational being, and a motivationally effective intellect more generally all can disregard them with impunity. An agent who has a motivationally but not fully effective intellect can, for practical purposes, ignore inclinations because regardless of their strength, reason outcompetes them in claiming the agent's attention. Conflicting desires and impulses are present, but outweighed. The claims of reason are simply stronger. We saw in Section 4.4 that what makes a motivationally effective intellect this causally efficacious is not merely the power of its occurrent psychological and neural states qua states. It is the added power of their rational content – and the recognition of oneself as rational in that content – which those thoughts and beliefs express that does the motivational work. In 4.5 we saw that an occurrent thought or belief may precipitate action because the rationality of its content prompts one to actualize one's rule-governed disposition to rationality in action that reflects this content. But in order to understand the activating impact of occurrently thought or believed rational content on a motivationally effective intellect – i.e. the causality of reason, as Kant would describe it, we need to situate this constellation of concepts in their broader social context. We can do this by revisiting a distinction in terms of which Chapter I framed this project, and that finds its origins in Nietzsche,[26] between two ideals of rational motivation.

6.1. Egocentric Rationality and the Ideal of Spontaneity

The first ideal of rational motivation is grounded in the value of spontaneity in action; in immediately expressing in one's behavior any impulse or desire that forms. An agent who embodies this ideal acts and reacts spontaneously to the extent that there is little or no delay between thought or impulse and action. For such an agent, to think it is to act it out, and to feel it is to express it in action. Because all thoughts and impulses are immediately vented in action, the agent's mental and emotional life is enacted in the external environment. And because of the rapidity with which all such thoughts and impulses are externalized, they do not linger in the mind long enough to develop pressure, weight, depth or mass. Aggressions as well as passions are relatively slight and transient disturbances in the agent's mental and emotional experience that become similarly slight and transient disturbances in the agent's external environment. Just as desire does not deepen into obsession or aggression into vendetta, similarly aversion does not deepen into hatred or passion into mania. Frustrations, irritations, hurts, fears, angers, even impulses to violence or vengeance are as quickly vented and easily released as are satisfactions, attractions, pleasures, thrills, or passions. The swift release of negative as well as positive thoughts and emotions into action provides an ongoing roller coaster of variegated highs and lows to which spontaneous agents immediately react with more of the same.

Were such reactive behavior to transgress settled social norms or violate a community's shared principles of self-control, it would not be tolerated for long. So such an agent can exist over time only in the company of other, similarly inclined agents – i.e. in a community in which spontaneity in action is itself a settled social norm. In such a community, there are bound to be many conflicts, battles and emotional outbursts, much bickering and remonstrating; fistfights, duels, and wars. But these, too, are relatively shallow disturbances in the agents' experience and environment that may either accumulate over time through reciprocal reaction, or, alternately, that may be just as quickly succeeded by peacemaking, pleasure, and mutual celebration – joyful events that are equally transient in their effects.

For in this community, the capacity for memory is not highly developed; there are few continuities of value or behavior (beyond the spontaneity of that behavior itself) to which individual events and actions are indexed. So grudges are not nursed unrelentingly and acts of heroism or mercy are not commemorated systematically or inscribed in tradition. Agents do not feel much guilt over past derelictions because these are difficult to recall and more difficult still to sanction. Nor do they derive much pride from past achievements, for much the same reasons.

Indeed, the very concept of achievement has restricted application in a community of spontaneous agents. Without a developed ability to index a succession of past events to a continuing history that extends into the present, spontaneous agents are similarly handicapped in their ability to project a succession of events into a continuing future that extends forward from the present. Of course they may have wishes, dreams, fantasies, and dissatisfactions they desire to be remedied. But these, too, are transient and volatile, regardless of their grandiosity, and evaporate or reconfigure with situational changes. Hence spontaneous agents have

comparatively restricted ability to plan or envision such an extended future beyond immediate satisfaction of present desire enacted in present action; and so are neither intimidated nor inspired by any such future goals or ambitions.

Consequently, impatience and boredom and an unending, reactive search for satiation and entertainment are among the few constants of behavior in a community of spontaneous agents. Because they have only minimal ability to locate themselves and their actions along a temporal continuum of such actions, and therefore minimal ability to evaluate those actions positively or negatively relative to settled and stable values, spontaneous agents tend to lack the metaethical preconditions for self-respect in the Rawlsian sense, of regarding their long-term goals and ambitions as worthy of achievement. Since spontaneous agents are limited in their ability to form long-term goals and ambitions, the question as to the worth of those goals and ambitions, or their ability to provide in a more sustained way the gratification spontaneous agents incessantly seek, generally does not arise. The funnel vision to which spontaneous agents are confined arises not from their preoccupation with satisfaction of long-term desires that saturates their perception of everything, as described in Volume I, Chapter II.2.3; but rather from the plethora of short-term desires that constantly demand attention.

Since its settled social norms are ill-suited to distribute scarce material resources justly or efficiently, a community of spontaneous agents such as this takes on an entirely different coloration, depending on whether its available material resources are abundant or scarce. In the latter case, their conflicts and battles over possessions and events are complicated and intensified by catastrophic instability: poverty, disease, malnutrition, homelessness, familial breakdown, social disintegration and criminality, which confine an agent's attention to immediate issues of physical survival and further discourage development of the capacities for foresight and long-term planning. These conditions are in turn further exacerbated by the imprudence and failures of impulse control that characterize spontaneous agents in the first place. Consequently, their pleasures and personal satisfactions are correspondingly immediate and often self-destructive. Spontaneity in the presence of severe material deprivation is a recipe for communal self-defeat, and so fails as an ideal of rational motivation.

So a community of spontaneous agents can survive and flourish only in an environment in which material resources are and continue to be abundant. In this case, conflicts among such agents are quickly forgotten when new attractions capture their attention. Similarly, abundant material resources must be presupposed by the limited ability of this community to engage in long-term planning. Without resources adequate to repair the damage caused by shortsightedness and to amend the repeated errors and blunders caused by forgetfulness, such missteps threaten to become fatal to its survival. So this ideal of spontaneity in action presupposes an environment in which resources are abundant enough to compensate for the mistakes and failures that come from imprudence and ignorance, and sufficient to insulate spontaneous agents from their harmful consequences. That is, they must be abundant enough to make any such failures seem insignificant – abundant enough, in effect, to instill carelessness as a

way of life. Spontaneous agents who have survived material deprivation with their spontaneity intact undergo no very profound transformation when supplied with material abundance.

Wealth of resources in conjunction with spontaneity of action thus lead this community of agents to be *self-anointing* in their value judgments. Because material abundance enables their lifestyles and empowers their choices, they naturally regard their unending supply of material abundance, and the power and status it confers on them, as confirming their inherent worth; as both motive and reward for the spontaneity of their behavior, and for the particular thoughts and desires they spontaneously express. Thus their wealth, their power and their position lead them to evaluate themselves, their actions and their circumstances – regardless of their soon-forgotten consequences – as intrinsically valuable; and others who are unlike them, conversely, as lacking in value; as insignificant and uninteresting afterthoughts. Although spontaneous agents lack self-respect, they do not lack self-worth or self-confidence. Their material abundance, power and status confer their sense of self-worth on them, confer authority and legitimacy on their actions, and so confer a sense of entitlement to perform them. Wealth, power and status make them who and what they are, and ensure their dominance. That is why Nietzsche describes such a community as one of *Übermenschen*.

A community of spontaneous agents (or, if you will, *Übermenschen*) is not capable of engendering a recognizable morality, i.e. a set of motivationally effective practices that establish and govern equitable relations between individuals with competing or conflicting interests. It is not capable of generating such a set of practices toward others who are unlike or outside that community, because spontaneous agents accord such others neither value nor influence in their affairs, and therefore have no incentive to engage with them. But such a community is equally incapable of engendering a recognizable morality to guide interactions even within that community itself. The pervasive preoccupation with self and immediate satisfaction is too overpowering, and the concern for long-term consequences too underdeveloped. Relationships within such a community, both personal and social, are a series of shortsighted and unregulated power struggles in which influence and status trump principle.

This ideal of spontaneity is recognizable as a variation on the limiting case of the utility-maximizing ideal described in Volume I, Chapter IV.4. We saw there that in this ideal case, the utility-maximizing agent achieves his ends instantaneously, with no expenditure whatsoever of time or energy – thereby achieving the smallest possible fractional proportion of resources expended to ends achieved. I also observed there that the limiting ideal of utility-maximization implies that the agent's adoption of an end physically and temporally coincides with its realization. It describes a situation in which the agent need perform no instrumental action at all. I argued that an ideally means-rational action is an atemporal set of instantaneous desire-satisfaction events, relative to which any instrumental effort at all is an unwelcome deferment of gratification. I called the utility-maximizing ideal *egocentric*, in that it describes a limiting case in which there is no order or sequence in which events should occur that is independent of when an agent desires their occurrence; in which the time, place, and manner in which desired events do

occur are entirely dependent upon the agent's desires as to when, where, and how they should occur. I concluded that such an ideal state is both worthless and meaningless.

The ideal of spontaneity as detailed above also illuminates the objection addressed in Volume I, Chapter III.2, that utility-maximization is not an end. Though we have seen that this objection is invalid for most empirical cases, it would seem to hold for the limiting ideal case. For spontaneous agents need not aim at efficiency, and may even lack the ability to formulate this long-term end. Because they need not "economize," they aim only at getting what they want. This ideal thus spells out some of the further psychological and social ramifications of the utility-maximizing ideal. It suggests the very limited extent to which the unreconstructed utility-maximizing model of rationality can provide an ideal of genuinely rational motivation at all.

During the mid-century development of the United States as a superpower and in the wake of the aggression, brutality and trauma of World War II, post-war Anglo-American analytic philosophy officially repudiated the existence of an inner mental life separate from overt behavior. Under the leadership of Wittgenstein, Ryle, Austin, Ayer and Quine, philosophy of language was dominated by behaviorism, an anti-psychologistic theory formulated by Watson in the years following World War I and elaborated by Skinner in the '40s. Behaviorism is the view that there is no "ghost in the machine," to quote Ryle; but merely a complex human machine that responds with attraction to positive reinforcement and with aversion to negative reinforcement. Behaviorism reduces all mental states and attitudes to behavior and dispositions to behave. We saw in Volume I, Chapter II.1.1 that Brandt and Kim's, and Lewis' analyses of desire sought, with mixed success, to exemplify this approach. Correspondingly, mind-body materialism prevailed in the philosophy of mind. In economics Samuelson's and Little's theory of revealed preference, that agents' preferences are revealed in their behavior, made its entrance at approximately the same time.

Anti-Rationalist metaethics is a refinement on this constellation of anti-psychologistic views. Of course Anti-Rationalism does not imply the immediate externalization in behavior of any desires, emotions, instincts, and impulses identifiable as such, as do its behavioristic antecedents. But it is in theory incapable of providing an account of how such inclinational mental events might be retained for long in consciousness, unexpressed, without coming under the purview of the intellect, and in particular those organizing functions of intellection that preserve horizontal and vertical consistency over time in the structure of the self. Hence even when such desires etc. power action in subsumed or sublimated form, reason is motivationally effective as a necessary condition and contributing cause. In order to be a sufficient condition and precipitating cause of action independent of reason, such desires must find immediate and spontaneous expression. In this form they maintain independence not only of reason, but therefore of rational intelligibility. So Anti-Rationalism's repudiation of reason and valorization of desire, emotion and instinct as a model of moral integrity constitute the "ethics" of the ideal of spontaneity – even as it calls into question whether the term can have meaningful application within that ideal.

The ideal of spontaneity and its associated constellation of values and practices thus have a long pedigree in Anglo-American analytic philosophy and Neo-Classical economics. Nevertheless, I regard it as anachronistic in its content and misguided in its philosophical implications. All of these theories – behaviorism, mind-body materialism, revealed preference theory, Anti-Rationalism – are variations or elaborations on the core anti-psychologistic ideal of spontaneity, that there is no significant, extended, interior life of the mind; that just as thoughts are externalized in language and exist only to that extent, similarly desires are externalized in action and exist only to that extent. If thoughts exist only in linguistic utterances and desires exist only in overt non-linguistic behavior, it is of course difficult to imagine how unuttered thoughts might motivate overt non-linguistic behavior. But I hope it is now clear that to accept the antecedent is unnecessarily restrictive.

6.2. Transpersonal Rationality and the Ideal of Interiority

A community of spontaneous agents of the sort just described obviously cannot be self-sustaining. Even an overabundance of material resources that serves no function other than to provide gratification and protection of desires and impulses cannot serve even that function without the patient ministrations that transform raw materials into tools, artifacts, and foodstuffs; and that maintain these props, implements, rewards and trophies of wealth in good working order. Spontaneous agents themselves by their very nature largely lack the capacities as well as the temperament necessary for managerial oversight as well as disciplined performance of these functions. These therefore must be assigned to a different community of agents altogether.

The second ideal of rational motivation is grounded in the value of interiority; of a vivid and extended life of the mind that includes imagination, intellection, and reflection; these are the foundations of transpersonal rationality. An agent who embodies this ideal suppresses immediate and untrammeled expression of thought or desire; withholds them from external view, drives them inward, transforms them into enduring mental and emotional presences, and forces an inward expansion in the purview of the mind's eye in order to accommodate them. The more fully such an agent internalizes thought and desire, the more intensely and vividly she experiences them and the greater her capacity for interiority becomes. Scope and depth of interiority is thus directly proportional to suppression and control of the impulse to immediately vent thoughts and desires in action. Thoughts and desires develop form, pressure, weight, depth and mass (i.e. horsepower, not to put too fine a point on it) as conative forces – sometimes emotionally explosive ones – within an agent to the extent that the agent resists their immediate externalization. Obsession, vendetta, hatred, mania and fanaticism are all distinct if extreme possibilities for thoughts and desires that find an outlet neither in spontaneous expression nor in reflective sublimation. Hence an interiorized agent must learn self-control: to express those thoughts and desires in the right way and at the right time.

At least at the outset, self-control cannot be voluntarily bootstrapped into a settled habit of character, as Aristotle sometimes seems to suggest. Rather, it is a reaction to external sanction,

in which spontaneous self-expression injures the agent – provokes punishment, threats, retaliation, or deprivation. It thus presupposes two conditions: first, spontaneous self-expression as an available model of behavior; and second, infliction of pain or negative reinforcement on an agent who attempts to emulate this model. So interiority develops within a community in which one's own spontaneity is discouraged by others' negative reaction to it, and self-control is a survival response to that negative reaction.

An interiorized agent therefore does not require a surrounding community of similarly interiorized agents in order to exist. Because his interiority is the product of self-control, and his self-control a survival response to punitive sanctions for spontaneous self-expression, he can survive in a community of spontaneous agents in which he is isolated from other interiorized agents. Because self-control grows out of a reaction to external exigencies, his interiority may thrive whether it is a settled social norm or not. In either case his interiority carries with it a certain degree of alienation, because he vividly experiences the contrast between the richness of his first-personal, inner experiences and the necessary superficiality of his third-personal observations of others' behavior. This contrast enhances his awareness of himself as a distinct and independent agent with a complex mental life at the same time that it complicates the facility with which he establishes deep connections with others. I discuss this contrast at greater length in the following chapter.

Such an agent is well suited to perform the managerial and labor necessities for a community of spontaneous agents whose luxurious lifestyle requires them but who are themselves temperamentally incapable of performing them. Nietzsche therefore describes the evaluative attitude that interiority engenders as a "slave morality" and interiorized agents themselves as *Untertanen*. However, when performed on behalf of a materially deprived community of spontaneous agents, this same evaluative attitude would be more aptly described as a supervisory, custodial or care-taking morality, in which the contrast in self-sufficiency and competence would track an opposing contrast in power and social status.

Interiority creates the internal mental and emotional space in which memory can flourish. This space is itself the enduring conscious presence relative to which experienced events and actions can be indexed. Because self-control requires that each such experience be inwardly retained rather than outwardly released, the accumulation of experiences over time elicits the sorting and classifying functions of intellection that rationally structure and systematize this interior space: discrimination, contrast, ascription, identification, subsumption, generalization, and so forth. These are the higher-order mental functions that Kant describes as "synthesis." These functions collect and organize the data of experience – including feelings, emotions, desires, and instincts – into a spatiotemporally continuous, rationally coherent self, and enable an interiorized agent not only to retain more than one item in mind at a time, but also to classify, recall and generalize over them systematically. Within the parameters of a rationally coherent, interiorized self, then, desire and emotion are brought under the purview of reason from the beginning.

With memory, imagination and intellection flourishes the ability to dwell on slights, nurse resentments, and exaggerate injuries – of which there are, of necessity, many, since the succession of these are what necessitate self-control in the first place. With these abilities also flourish the abilities to recall and reflect on the past, to derive meaning, illumination and satisfaction from it; to extend the lessons and values learned from it into principles and theories that guide present action and future planning; and to imagine counterfactual alternatives to actual states of affairs. In the flowering of these abilities consists the development and growth of transpersonal rationality. They extend the agent's awareness past the boundaries of the present moment and situation, past the boundaries of the body and the self, past their felt needs and drives, past the boundaries of an individual human lifespan, and indeed far past the boundaries of the physical world. The achievement of interiority releases individual awareness from the funnel vision of immediate drives and impulses into an unbounded universe of theory-laden modality, in which necessities, facts, and possibilities at all levels of abstraction compete for the agent's interest. Interiority at this level is an authentic expression of transpersonal rationality.

The more complexly, vividly and clearly the agent renders this universe in her mind's eye, the more it compels her attention and draws it away from the confining personal drives, desires, and needs whose deprivation supplies its impetus; and the more the pull of those drives themselves recede into the background of awareness. Then the easier it becomes to contemplate envisioned objects, scenarios and principles impersonally, without regard to their particular relation to the agent herself. That is, the easier it becomes to dream and take flight beyond the limits of the self. Abdicating the primacy of the agent's own desires and impulses to the demands of her interior universe further enhances the form, weight, depth and power of this by now quite vast expanse of interiority; and marks the juncture at which transpersonal rationality may outcompete the egocentric drives of the ego itself as a conative force. The modal operations of imagination engender images and visions of as yet-unrealized alternatives, while those of the intellect refine, sharpen and systematize their details. These find form in works of art, architecture, literature, music, philosophy, and in religious, spiritual, and scientific inquiry – in fact, in all of the intellective practices and artifacts of a civilization that transcend the personal and physical limitations of individual agents. Transpersonal rationality of this kind is not only compatible with creativity but also a precondition of it.

Spontaneity denied is action and satisfaction deferred, projected into a future indexed by subjectively probable events uniformly with those which have been recalled from the past. Memory and intellect thus force interiority to expand even further, to encompass visions, goals and ambitions that gain substance and power from the complex and painful psychological operations that engender them. With these come fantasies of victory, revenge and retribution for wrongdoing – as well as corrective ideals of justice, goodness, beauty and truth and their corresponding emotional reactions. The capacities for prudence and for altruism develop jointly and simultaneously. By their nature, then, these corrective ideals as well as their shadow fantasies thrive in conflict with external realities of hardship, servitude, deprivation or injustice.

Indeed, agents who make a conscious and deliberate commitment to a life of interiority, i.e. of the spirit, sometimes practice voluntarily self-enforced disciplines and austerities that are, in effect, specialized techniques of deprivation, hardship or servitude – for example, fasting, celibacy, vows of poverty, renunciation, obedience or service – as means to accelerate and deepen interiority and the insights and wisdom it offers.

As these gain in intensity, vividness and clarity, they compete increasingly with the external realities on which they improve. That is, they function not only as alternatives to external reality, but therefore as powerful criteria against which those external realities are judged. Hardship in conjunction with self-control thus lead interiorized agents to be *other-anointing* in their value judgments. Comparing external events, circumstances and other agents with the developed interior universes they carry around inside them, they pass judgment on those exterior conditions with an eye to their violation of or conformity to the interior orders they have created. Whereas external conditions – of material abundance, status and power – induce spontaneous agents to confer value on themselves, interiorized agents confer value on or withhold value from the harsh and imperfect external conditions they contemplate. Transpersonal rationality may not be married to any particular moral theory; but it is inherently evaluative, and therefore critical.

Thus interiorized agents of necessity develop the ethical capacities for impersonality, disinterest, selflessness, and impartiality that are engendered by the pleasures of abstract speculation and inquiry, in direct proportion to the vividness, clarity and power of the interior universes they are forced by circumstance to create. Alongside these ethical capacities, interiorized agents also develop the unethical capacities for calculated revenge, betrayal, deception, and self-aggrandizement. When exercised, the ethical capacities in turn nourish imaginative insight into others' inner states, and enable the empathic moral emotions that such insight calls forth. Hence just as – as we saw in Chapter IV.8 – memory provides the intrapersonal foundation for the negative moral emotions of guilt, shame and resentment, similarly imagination provides the interpersonal foundation for the positive moral emotions of empathy, sympathy, pity and compassion. I offer an account of the intrinsic interconnections among imagination, impartiality and compassion in the following chapter.

The most effective way to quash corrective ethical ideals and their unethical shadow fantasies, or to soften their hard edges, is to amply redress those external realities – in effect, to buy and entitle through the bestowal of material abundance the agent who entertains them. This strategy has two results. First, it gradually reconditions an interiorized agent into a spontaneous one, by arrogating to her desires and instincts the power, authority, legitimacy and freedom of spontaneous self-expression. This returns the agent to the immediate fact of her concrete physicality, by indulging, legitimizing and valorizing her "gut impulses." Second, therefore, it gradually disconnects her from the infinitely expansive interiority of the intellect and imagination. By contrast with spontaneous agents, then, interiorized agents who have survived material deprivation undergo a quite remarkable transformation when supplied with material

abundance. In general, material deprivation tends to transform spontaneous agents into interiorized ones, and material abundance tends to transform interiorized agents into spontaneous ones.

The weight, complexity and motivational power of fully developed transpersonal rationality are largely inaccessible and inexplicable to spontaneous agents. Because the psychological lives of the latter occur primarily in the external physical environment – in overt action, reaction, and their consequences, the very idea of a highly developed interior, nonphysical order of psychologically continuous and rationally organized events is difficult for a spontaneous agent to comprehend. The idea that such an order should engender considerations that outweigh the importance, interest or urgency of an agent's own spontaneous and immediate desire-satisfaction is doubly unintelligible. Most unintelligible of all is the suggestion that such an order, and the transpersonal visions and principles it engenders, might attract greater interest, attention, or behavioral response from the agent whose interior order it is than the gratification of that agent's egocentric needs and desires.

But to an interiorized agent, there is nothing so remarkable about this. An interiorized agent may be in the grip of an impersonal, interior vision of justice, goodness, beauty or truth in the same manner in which a spontaneous agent is in the grip of his personal desires. An interiorized agent may be propelled into action by ethical or spiritual conviction, or by compassion or moral outrage, in the same manner in which a spontaneous agent is propelled into action by his drives or desires. The difference between them is that the spontaneous agent makes a singular judgment about the goodness of all of his drives and desires, whereas the interiorized agent makes a comparative judgment about the superiority of her capacity for interiority itself. She is causally influenced by the impartial directives it engenders because these define and make intelligible to her the kind of self she is. Interiority, then, is a necessary condition of transpersonal rationality; and this, in turn, is a necessary condition for the development of a recognizable morality in the sense defined above. Even a morality fashioned for the sole purpose of coordinating conflicts among mutually disinterested individuals requires more of those individuals than mere attention to the instruments by which they realize those interests. It requires in addition the complex functions of memory, imagination, intellection and reflection just described; and the civilizing moral emotions that develop in response to them.

The foregoing ideal of interiority is recognizable as an elaboration on the ideal of vertically consistency described in Chapter II, above. There I argued that the highest-order concept of the self-consciousness property, i.e. the concept of being an object of experience one has, subsumes as a matter of conceptual necessity all lower-order things and events that are rationally intelligible to one at a given moment and that therefore constitute one's perspective; and that we must be able to make everything, including objects of our experience, and objects of objects of our experience, and so on, rationally intelligible to ourselves as objects of our experience – i.e. in terms of the highest-order concept of the self-consciousness property, in order to be genuine agents at all. An agent's perspective as thus defined is a self-reflective, interior

perspective on external things and events that is saturated and fashioned by her transpersonally rational understanding of them. A subject for whom all such lower-order things and events constitutive of her perspective are rationally intelligible, and who therefore has the capacity for agency, therefore has the capacity to act *or to refrain from acting* in response to her transpersonally rational perspective on those external things and events. She self-reflectively controls her external behavior in response to an interior comprehension of what her external circumstances require.

As we have seen, interiority is not the same as fully effective intellection independent of desire or inclination; it encompasses all of these and more. Desire and inclination deferred aid in fashioning interiority. Through their sublimation and subsumption by the intellect, they become rationally intelligible elements within a self unified by the horizontal and vertical consistency over time of all its experiences and defined, in part, by the particular desires and inclinational experiences it has. But if self-control, and the rationally intelligible interior perspective it fosters, are necessary conditions of genuine agency, then only the ideal of interiority can explain how rational principles such as 4.4.2.(6) – (8), and the moral principles detailed in Section 5.2 above, can precipitate an agent into action. They can do so because interiority is structured and shaped by the sorting functions of intellection. We have seen how these have, of necessity, a governing and – in the ideal case – sufficient role in initiating action. Even in the non-ideal case in which delinquent desires and impulses are strong, highly developed interiority offers an agent a transpersonally rational perspective whose principles habituate a disposition to rational action that is stronger.

If the foregoing elaboration of this ideal enables us to imagine how unuttered thoughts might motivate overt non-linguistic behavior, then it is unlikely that, as behaviorism, revealed preference theory and mind-body materialism stipulate, thoughts exist only in linguistic utterances and desires exist only in overt non-linguistic behavior. Therefore, we need the ideal of interiority in order to understand how substantive moral theorizing can inspire and motivate the implementation of alternative agendas, disinterestedly and independently of the immediate self-interests and biases of their proponents. The following analysis of compassion in terms of impartiality and modal imagination demonstrates how the complex workings of interiority might move a transpersonally rational agent to balanced and disinterested moral action independent of personal desire.

Endnotes to Chapter V

[1] The term is Elijah Millgram's; see his "Does the Categorical Imperative Give Rise to a Contradiction in the Will?" *The Philosophical Review 112*, 4 (October 2003), 525 – 560. Millgram does not discuss Baron's analysis, because his target is recent Kantian accounts of rational deliberation, rather than of rational motivation. But the family resemblance of Baron's approach to the accounts he does discuss, as well as the explicit influences she cites, warrant the inclusion of her analysis under the same rubric.

[2]This is a common phenomenon among cancer patients. An example of the psychological restructuring of the self in response to painful assault of a more adulterated sort can be found in discussions of Post-Traumatic Stress Disorder. See *DSM-III: Diagnostic and Statistical Manual of Mental Disorders*, Third Edition (Washington, D. C.: American Psychiatric Association, 1980), 236-9.

[3] See Barbara Herman, "On the Value of Acting from the Motive of Duty," *Philosophical Review 66* (1981): 359-382; reprinted in her *Practice of Moral Judgment* (Cambridge, Mass.: Harvard University Press, 1993).

[4] Marcia Baron, *Kantian Ethics Almost without Apology* (Ithaca: Cornell University Press, 1995), 113. Henceforth all references to this work are paginated in the text.

[5] In "Autonomy: The Emperor's New Clothes," The Inaugural Address, *Proceedings of the Aristotelian Society, Supp. Vol. LXXVII* (2003), 1-21, Onora O'Neill contends that Kant means to characterize maxims as determinations of the will in the sense of being a formal, not an efficient cause (8); but the language Kant uses in many of the passages I list here suggest, rather, that he thinks of the form of law itself as an efficient cause. O'Neill justifies her reading of the texts on the grounds that "[t]he principle (law, rule plan) that an agent adopts does not cause him or her to do anything (how could abstract entities such as *principles* (*laws, rules* or *plans*) be efficient causes?)." I take my task in this chapter to be to answer that question. The nutshell answer would be that such a principle can be an efficient cause by being the rational content of an occurrent thought- or belief-event in which a rational agent recognizes herself, such that this self-recognition in turn efficiently causes her action.

[6] Henry Allison, *Kant's Theory of Freedom* (Cambridge: Cambridge University Press, 1990).

[7] I do not mean to identify Henry Allison as a New Kantian and am fairly sure he would not identify himself as one.

[8] Onora Nell [née O'Neill], *Acting on Principle: An Essay in Kantian Ethics* (New York: Columbia University Press, 1975), esp. Chapter Five.

[9] I offer an analysis of maxims in "Kant on the Objectivity of the Moral Law," in Andrews Reath, Barbara Herman and Christine M. Korsgaard, *Reclaiming the History of Ethics: Essays for John Rawls* (New York: Cambridge University Press, 1997), 240-269.

[10] *Op. cit.* Note 8, 77. Also see O'Neill's later account of conceptual inconsistency within a maxim in her "Consistency in Action" (*Constructions of Reason* (Cambridge: Cambridge University Press, 1989), 89). Both (1) and (2) could also be read as failing O'Neill's Fourth Principle of Rational Intending, that "the various specific intentions we actually adopt in acting on a given maxim in a certain context be mutually consistent" ("Consistency in Action," 92), *if* (1) and (2) are interpreted as implicitly containing multiple intentions that conflict. However, both are phrased as singular intentions.

[11] I am not convinced that this is the kind of intentional object that *can* be ascribed to an intention of the ordinary kind, for the reasons discussed in Chapter II.2 above, but leave that aside for now.

[12] And in the following analysis I anticipate and appropriate to my own account a principle of exegesis I apply to Kant throughout *Kant's Metaethics*: that if Kant develops in the *Critique of Pure Reason* an extended, technical use for a term (such as "reason," or "recognition," or "idea," or "categorical," for example), it is not a good idea to ignore that technical use when the same term occurs in later works such as the *Groundwork* or second *Critique* that seem clearly intended to build upon its foundations.

[13] Of course thus recognizing oneself is an experience one has, and therefore not higher in order than the self-consciousness property.

[14] Besides, I like it.

[15] I defend this interpretation of Kant at greater length in *Kant's Metaethics*. The argument is summarized in outline in "Kant on the Objectivity of the Moral Law," *op. cit.* Note 9.

[16]*Nicomachean Ethics*, trans. Terence Irwin (Indianapolis: Hackett, 1985), 1105b8.

[17]*ibid*., 1113a30.

[18]Henry Sidgwick, *The Methods of Ethics* (New York: Dover, 1966), p. 33.

[19]*ibid*.

[20] 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.

[21]One should not be misled by the singular subject of each of these propositions. But for ease of exposition, each occurrence of "a rational being" could be replaced by "any rational being." Hence each of these statements takes a universal quantifier.

[22]Here and below I make an unargued assumption about the semantic equivalence of Kant's use of the terms "noumenal" and "intelligible" on the one hand, and "phenomenal" and "sensible" on the other. To defend this assumption would require a paper of its own, but I have every faith that such a defense would succeed.

[23](A.4) is added as part of a fuller explanation of parents' behavior than that to which a child ordinarily has access. From the child's perspective, that its parents made a promise may seem a sufficient explanation of their keeping it. But in fact, its parents' perceptions that the child

needs to experience exemplars of promise-keeping, to have its self-esteem nurtured by parental support, etc., may also play an explanatory role.

[24]My choice of the term "construct" over "entity" should not be taken to imply a commitment to operationalism. I use it because it sounds ontologically peculiar to describe many of the higher-level concepts of Theory *K*, for example, "reason," "law," "humanity," "freedom," etc. as denoting entities.

[25]in the sense that Weber defines in *The Theory of Social and Economic Organization*, Ed. Talcott Parsons (New York: Free Press, 1964), Chapter I.1 and I.2.

[26] Specifically, the first and third essays in *On the Genealogy of Morals* (Friedrich Nietzsche, *On the Genealogy of Morals*, trans. Walter Kaufmann and R. J. Hollingdale (New York: Vintage Books, 1967), Essay I: “Good and Evil,” “Good and Bad”; and Essay III: “What Is the Meaning of Ascetic Ideals?”). However, although my analysis in this and the following section is inspired by Nietzsche, it is not intended as Nietzsche exegesis.

Chapter VI. Moral Interiority

In Chapter V I sought to show how reason, as formally defined in Chapters II through IV, could provide both necessary and sufficient motivational conditions for action. I emphasized that whether screening genuine preferences in the first case, or directly precipitating action in the second, any such motive that meets the requirements of a genuine preference, including desires and emotions, qualifies as rational on this account. Because I felt it necessary to address first and foremost the issue of how an occurrent thought or belief could both precipitate and guide the action it motivates, I did not say much about how other kinds of motives, particularly affective ones, might fit into the schema of a genuine preference. I now attempt to remedy that lack.

I argued in Chapter V.6.2 that interiority was the key to understanding how the capacities constitutive of transpersonal rationality could have causal force. I also contended that interiority is not a purely intellective condition of the agent, but rather one that subordinates all intelligible experience, including emotions and desires, to the sublimating functions of the intellect. With the aid of the concept of interiority as an ideal of rational motivation, I now turn to a more detailed look at how interiority, and in particular our intellective capacity for modal imagination enhances the conative power of a motivationally effective intellect, refined and suffused by moral emotion, to move an agent to specifically moral action, again independently of desire. I defend this thesis, first, by offering a conceptual analysis of compassion – a complex, transpersonal moral emotion that exhibits the effects of such sublimation but is not easily reducible to desire. Although compassion is itself a normative moral concept, nothing I say here carries any particular prescriptive commitment as to the relatively central or peripheral role I might think compassion should play in a normative moral theory. So, for example, the analysis that follows is consistent with a moral theory that advocates the motivational priority of moral duty (or, for that matter, personal loyalty) over compassion when the two conflict. I develop metaethical criteria that constrain the choice of an adequate normative moral theory in Chapter X below.

Compassion is not the only moral emotion. But it is a paradigmatically transpersonal one, in part precisely because it involves no desire (in the nontrivial sense) but nevertheless has both broad transpersonal scope and considerable conative force. On the following analysis, compassion involves both intellective and affective capacities: first-/third-person symmetrical modal imagination, empathy, sympathy, a disposition to render aid or mercy, and what I describe as strict impartiality. It therefore involves several of the key capacities of interiority discussed in the preceding chapter. My main aim here is to demonstrate how these capacities saturate an agent's perspective with an ideal descriptive moral theory of the sort described in Chapter V.5.2 above; thereby extend the agent's awareness and attention beyond the limiting constraints of egocentric desire; and how this very transpersonal extension of concern itself increases the conative resources available to a motivationally effective intellect.

In Section 1 I define the key concepts of modal imagination and impartiality I use as tools in the analysis to come. I distinguish inherently impartial principles from those which may or may not be applied impartially. I am concerned with the latter. I then distinguish between impartiality and impersonality (here reprising some material discussed in Volume I, Chapter VIII.3.2), and between impartial judgment and impartial treatment. I contrast my understanding with Lawrence Blum's, whose views I take up in greater depth in Section 5. Section 2 is devoted to spelling out the implications for our relations with others of our ability to imagine modally their interiority. I argue that without it our conceptions of ourselves and others would be primitively egocentric and narrowly concrete, in ways that would impede the simplest acts of interpersonal understanding or coordination.

Section 3 distinguishes two ways in which we envision objects of modal imagination: as surface and as depth objects; and two extremes relative to which we exercise our capacity for modal imagination: self-absorption and vicarious possession. These two distinctions cut across each other. From them I derive two criteria for an appropriate level of involvement in a modally imagined object. Section 4 applies these criteria to an analysis of compassion as including empathy, sympathy, and satisfaction of a symmetry requirement in one's compassionate response to another's suffering. Section 5 brings this apparatus to bear on my objections to Blum's argument against impartiality, and introduces my contrasting notion of strict impartiality. Section 6 sharpens this concept of strict impartiality and shows how it resolves the purported conflicts between impartiality and compassion on which Blum insists. Section 7 revisits the twin problems of moral alienation and moral motivation discussed in Volume I in light of this analysis of compassion, and argues that on a Kantian conception of the self, these problems disappear. Section 8 closes this chapter by invoking my analysis in answer to the question left hanging in Volume I, Chapter VI.5.2, as to how to explain the motivation of a whistle-blower in the absence of Humean assumptions about motivation and rationality.

1. Impartiality

I begin by sharpening the term "modal imagination," introduced in passing in the analysis of interiority in the preceding chapter. I use it to call attention to a specific feature of imagination as we ordinarily conceive it. This is that we can imagine not only what actually exists, such as the computer screen now in front of me; but also what might have existed in the present or past, or might someday exist in the future, such as a vintage restored1950 Remington Rand typewriter. The term *modal imagination* is intended to remind us of our interiorized capacity to envision what is counterfactually possible in addition to what is actual. I argued in the preceding chapter that modal imagination is what extends our conception of reality – and in particular of human beings – beyond our immediate experience in the indexical present. Here I argue further that we need to do this in order to preserve the significance of human interaction. To make this leap of imagination successfully is to achieve, not only insight, but also an impartial

perspective on our own and others' inner states that recognizes and respects the symmetry between them. This perspective is a necessary condition of experiencing compassion for others.

My conceptual analysis of compassion depends on the key concept of strict impartiality. In Section 6 below, I show that strict impartiality differs from impartiality in the ordinary sense, by adhering more closely than impartiality in the ordinary sense to the spirit as well as to the letter of what impartiality in the ordinary sense explicitly requires. However, I also show strict impartiality to be similar to impartiality in the ordinary sense, in that both are metaethical requirements on normative moral principles of judgment and conduct, rather than normative moral principles themselves. So my later analysis of strict impartiality requires establishing here what impartiality in the ordinary sense comes to.

In the ordinary sense, a substantive principle is *inherently* impartial if it contains no proper names or rigged definite descriptions. But an inherently impartial principle may be *applied* prejudicially if it is applied only in some relevant circumstances and not others, or applied to suit the interests of some individuals and not others, or applied on the basis of attributes irrelevant to those explicitly picked out by the principle. So, for example, I violate the metaethical requirement of impartiality if I apply the principle of hiring the most competent candidate for the job only to the pool of candidates selected from a particular club or class or gender or race. This applicative notion of impartiality is also part of the ordinary usage of the concept. I am concerned with impartiality in this sense, in which it is the application rather than the formulation of the principle that is at issue.

In the applicative sense, to be impartial in one's *judgment* is to ascribe an evaluative predicate to a subject on the basis of the attribute or attributes the predicate denotes rather than on the basis of some other, irrelevant attribute which one happens to value or disvalue. Without knowing what the substantive judgment is and on what attributes it is based, there is no way of determining whether or not one has judged impartially. For example, my judgment that you would make a particularly entertaining dinner guest is impartial if it is based on the high quality of your conversation and social skills, and biased if it is based on your impressive professional connections. Without knowing what it is I am judging and on what attributive basis, whether or not my judgment is impartial cannot be determined.

Note that the impartiality of my judgment has nothing to do with whether or not I bear some personal relation to you, i.e. with how *impersonal* I am in making the judgment. Thus basing my judgment of your suitability as a dinner guest on your professional connections does not require that I be in the process of considering whether to invite you to dinner; or if I am, that I desire access to your impressive professional connections. There is nothing about failing to stand in personal relation to you that ensures impartiality of judgment, and nothing about standing in such relation that precludes it. Of course this is not to deny that that standing in a certain kind of personal relationship to you may tempt me to bias the application of my substantive principle in your favor, for instance if I want to curry your favor or avoid incurring your wrath. But this is just to acknowledge that impersonality, which is a psychological state, may, under certain

circumstances, facilitate adherence to impartiality, which is a cognitive norm. It is not to conflate the two, and – as I argued in Volume I, Chapter VIII.3.2 – there is no psychological reason to suppose that they must always go hand in hand.

Similarly, to *treat* others impartially is to be guided consistently in one's behavior toward them by an inherently impartial, normative principle of conduct, such that one acts as the principle prescribes and in accordance with the attributes its evaluative predicates denote, and not in accordance with other, irrelevant attributes one happens to value or disvalue. Again, without knowing what the substantive principle of conduct is, and on what attributive basis I am applying it, there is no way of determining whether or not my treatment of the other is impartial. So, for example, you cannot know whether I have treated you impartially in hiring you for the job unless you know, first, that my choice is guided by the principle of hiring the most competent candidate for the job, and second, that I have hired you because of your competence and not because of your club, class, gender or race. I am concerned with impartiality in this latter sense, in which it is the application of inherently impartial principles of conduct (rather than principles of judgment) that is at issue. I argue that compassion is a substantive moral emotion that disposes one to apply the normative principle of rendering aid to the needy and satisfies the metaethical requirement of strict impartiality as I define it below.

Lawrence Blum's view of impartiality differs from mine with respect to at least two of these claims. First, Blum criticizes Kantian moral theories on the grounds that in assigning a major role to impartiality, they thereby "deny a substantial role to sympathy, compassion, and concern in morality and moral motivation."[1] Although Blum does not define what he means by "compassion," he does say about impartiality that it involves "not giving weight to one's own preferences and interests simply because they are one's own, but rather giving equal weight to the interests of all, ... favoring none simply because of personal preference[.]" (44) Impartiality, on Blum's conception, is not an appropriate requirement where friendship is concerned. (46-66) My argument below implies that, like compassion, genuine friendship – as opposed to excessive dependency or insensitivity – would be impossible without it.

Second, Blum's characterization of impartiality as "giving equal weight to the interests of all, ... favoring none simply because of personal preference" does not clearly identify impartiality as a metaethical rather than a normative moral principle. It thus leaves open the conceptual possibility of normative *pseudoimpartialist* principles which might, for example, require one to treat everyone with a similar degree of detachment, or to distribute resources in exactly equal amounts to everyone, or to ascribe to everyone, including oneself, exactly the same predicates, all regardless of attributive basis. These principles would prescribe a policy, not of impartiality, but of indiscriminacy. As normative moral principles they would be very peculiar, and I know of no philosopher who holds any of them. They would also violate the metaethical principles of impartiality in judgment and treatment earlier described, since the indiscriminacy of their application would be inherently biased against certain cases identifiably demanding of special consideration by virtue of circumstance.

Blum himself does not explicitly describe the target of his criticism in normative pseudoimpartialist terms. But he does contrast what he thinks impartiality requires with what he thinks compassion requires with respect to actual moral conduct. Since compassion is a substantive moral concept, this contrast suggests that he views impartiality as a substantive, normative moral concept as well. I find this interpretation implausible for the reasons just mentioned. So I assume in what follows that we both mean to address the concept of impartiality as a metaethical criterion for the correct application of normative moral principles.

2. Modal Imagination

Begin by considering what our conception of human beings would be like without the modal aspect of imagination that extends our interiority beyond the actual. We would be able to recollect experiences and emotions we had had, as well as mentally envisage objects, events, and states of affairs we were presently experiencing. Images of familiar human bodies, stationary and in motion, silent and audible, as well as some of our intellectual, psychological, and sensory reactions to them, and our present reactions to those, would be among the items accessible to memory and visualization. Our conception of human beings would consist, roughly, in our sensory experience of ourselves and other human bodies, plus our complex reactions to them. We might experience cravings, needs, desires, and intentions in ourselves. But we could envisage neither absent objects of desire, nor ourselves satisfying those desires, since this would require us to imagine a possibility of action that we had not yet experienced (of course this is not to deny that we might in fact satisfy them nevertheless). A nonmodal conception of interiority, then, would be one in which our intentional states were experienced as events without foreseeable consequences.

Nor could we envisage other people satisfying their cravings, needs, desires or intentions, for the same reason. In fact, we could not imagine other people *having* these or any of the other inner experiences that constitute our interiority. Thoughts, emotions, desires, and sensory responses would constitute part of our conception of ourselves, but not part of our conception of others. Since each of us can experience only our own responses and not someone else's, and since we could imagine only what we had experienced, others' experience would not be accessible to our imagination at all.

Without the capacity to envisage events or states of affairs other than those we ourselves were experiencing or had experienced, we would be unable to identify our experiences in terms of universally applicable concepts, concepts that apply equally well to classes of events that may occur in the future or might have occurred in the past, in addition to those that are occurring in the present or did occur in the past. This means that, in particular, the concepts in terms of which we understood even our own inner states would be extremely limited. For example, no quantity of recurrences of certain kinds of emotional state would be sufficient to lead us to formulate the concept of love, or fear, or anger, or joy as we actually understand those concepts, because the application of each extends past the experiences we have actually had forward into a possible

future, and backward into a counterfactually possible past. So not only would others' inner states be imaginatively inaccessible to us. Our insight into our own would be almost nonexistent, or at least extremely primitive. We would experience our inner states as we do subtle changes in the weather for which we have no words.

Without the concepts that denote at least our own inner states, our capacity to reason about them or others' – to draw analogies, inferences, and conclusions, or to make inductive empirical generalizations about them – would be correspondingly reduced. For example, we might be able to juxtapose two or more experiences we had had, and perhaps even note the differences and similarities among them. But we could supply no term to any analogy that required us to posit an experience that was in some respect unlike any we had had. So in particular, I could not draw any analogy between any experience I had had and one you might have. Because your having an experience is not itself an experience I would have had, I would have no basis on which to conceive the possibility of your having an experience at all. Thus I might experience the piano landing on my toe, resultant shooting pains in my toe, and myself jumping up and down holding my foot, the surrounding visual horizon rising and falling accordingly. But from my observation of the piano landing on *your* toe and *your* jumping up and down holding your foot, I would fail to supply the corresponding sensations of the piano's landing on your toe, the resultant shooting pains, or your jumping up and down. Because I experienced my own behavior entirely first-personally and yours entirely third-personally, I would be unable to detect the relevant similarities between my behavior and yours. I would lack the imaginative basis on which to make even the simplest inference from the one to the other.

The result would be a primitively egocentric and narrowly concrete conception of human beings, in which the most vivid and memorable events were intrinsically tied to our sensory experience of others as mobile physical beings, and our intellectual and emotional responses to it and them. This conception would be *primitively egocentric* in that the criterion of significance in evaluating and judging our own and others' behavior would be some function of our own visceral response to them: the psychological quality of our reaction, for example; or its degree of pleasantness or vividness; or the ability of that behavior to arrest our attention. A primitively egocentric conception of others is not necessarily a selfish conception of them, since it does not necessarily evaluate and judge others' behavior with respect to the satisfaction of one's own needs and interests. A primitively egocentric conception is rather one that evaluates and judges another's behavior in accordance with the centrality of one's own experience: other people are more or less important or valuable, and their behavior more or less interesting or worthy of note, in so far as they viscerally move one – in whatever direction – to a greater or lesser degree. A primitively egocentric conception of others reverses the psychologically and morally intuitive order of events in moral appraisal: Ordinarily it is supposed that we are moved by an event or action or state of affairs because it is significant. An agent who holds a primitively egocentric conception of others regards an event or action or state of affairs as significant because she is moved by it.

The conception of human beings that resulted from a nonmodal imagination would also be *narrowly concrete* in that our view of ourselves and others would be neither informed nor inflamed by implicit, tentative suppositions regarding our or their internal motivations, thoughts, or emotional states; by hopes or expectations about our or their future behavior; or by speculations on possible courses of action revealed by our or their present behavior. We can assume for the sake of argument that our own motives, thoughts, and emotional states would be experientially accessible to us in some conceptually limited way, perhaps as schematic conjunctions of images.[2] But we would lack the capacity to speculate on the conceptual identity of those states in ourselves, just as we would lack the capacity to conceive them as being of any sort at all in others. Nor could we plan for the future, aspire to achieve goals, or consider alternative courses of action we might take. Our mental lives would be restricted to experiencing our present inner states and remembering past ones, and observing others' behavior and reacting to its impact on us.

Our social relations would be correspondingly bereft. Communications about plans, hopes, dreams, or desires would be nonexistent, as would the corresponding dimensions of personal character these intentional states express. The very ideas of sharing one's thoughts, reaching agreement, or achieving understanding with another would be unintelligible. Such relations might be somewhat more vivid to sensation without the intervention of suppositions and expectations about the other. But they would also be harsher, bleaker, and inchoate. They would lack the significance and depth conferred by our implicit presumption of potential. They would lack the richness of mutual insight conferred by shared emotions and thoughts. And there would be no place in such relationships for the mutual contentment and familiarity borne of a common worldview or value commitment, nor for the cooperative behavior that makes them possible.

Many of us have occasionally experienced primitively egocentric and narrowly concrete relationships, whether as object or as subject. Ordinarily we think of them as unsatisfactory and without future, and we try to improve or move past them. In the scenario I have been envisioning, in which modal imagination of alternative possibilities is foreclosed, even the conceptual possibility of moving past such "dead-end" relationships would be foreclosed as well. Virtually our entire ability to think about and understand our experience, both of ourselves and of others, as well as our ability to coordinate our behavior with others, presupposes the functioning of modal imagination. Those inclined to Cartesian scepticism about the existence of other minds need to be reminded of the centrality of modal imagination to the functioning of human social and mental life. And their verificationist fears need to be met with a reminder of what that life would be like without it.

3. Self-Absorption and Vicarious Possession

Next consider two extremes of imaginative object. At one end of the spectrum, there is the kind one effortlessly calls to mind with no cue beyond that of a momentary association or

verbal description. For example, I now ask you to imagine yourself rising from your seat, flapping your arms vigorously, and sailing aloft. It probably does not require very much mental concentration for you to activate the required visual imagery and subliminal sensations; the mere verbal description may suffice. However, easy come, easy go. Virtually any actual internal or external cue will suffice to banish that fantasy: the ringing of the telephone, your shifting in your chair, or something you read here that momentarily catches your attention. Call this a *surface object* of imagination. At the other end of the spectrum, *depth objects* of imagination call forth a deeper psychological investment of energy and attention. They occupy a larger proportion of one's waking consciousness, and may either replace or vividly enhance reality as one experiences it. For example, I read a first-person account by a battered wife of her experiences, and my emotions as well as my thoughts are fully engaged, not only as I am reading, but afterwards as well. My imaginative reconstruction replaces reality as I am absorbing her story, and alters my view of the world afterwards. Whereas as surface objects of imagination barely affect the quality of one's interiority, depth objects shape it profoundly in ways that may permanently alter one's perspective. Most imaginative objects lie somewhere between these two.

Clearly this taxonomy of imaginative objects is far from exhaustive. Nor does it sort imaginative objects into those we visualize and those we conceive in some more abstract or schematic sense: I may be deeply involved in imagining the outlines of my cosmological theory of the universe, or only momentarily distracted by the visual image of the groceries I must purchase on the way home. Whereas nonmodal imagination precludes imaginative conceptualization, modal imagination, as already suggested, supplements rationality to produce it.

Nor does the distinction between depth and surface objects of imagination classify such objects by content: Penrod Schofield was so fully engaged by the first-described fantasy that even Miss Spence's repeated shouting scarcely sufficed to return him to the reality of the classroom. Rather, I mean to distinguish among such objects of imagination according to the degree of one's momentary experiential involvement in them. Some such objects hold us in their grip, while others slide over the surface of our interiority while barely disrupting our emotional and psychological awareness at all.

Sometimes we treat as objects of surface imagination those we are called upon to treat in depth. For example, charitable concerns often bulk-mail letters to potential contributors that describe in vivid detail the plight of those for whom they wish to garner support. Upon receiving these mailings, one skims the letter, barely registering the import of the words, before tossing it in the trash. Conversely, we may treat in depth imaginative objects that are more deserving of surface treatment. For example, one may die a thousand deaths imagining in excruciating detail the possibility that one may flub a line the next time one presents a paper. The vividness of this scenario may overwhelm one with such serious anxiety or depression that it interferes with one's sleep patterns. In both of these cases, something has gone awry. In the first, one's level of imaginative involvement is, at least on the face of it, insufficiently responsive to another person's

real crisis – a predicament that demands a considered and fully attentive response to it. In the second case, one's level of imaginative involvement is excessively responsive to an inconsequential possibility that can be prevented easily (for example by rehearsing a few times beforehand one's delivery of the paper).

Naturally, each of these inappropriate imaginative responses could be directed towards the other imaginative object. It may be, for example, that one is so engaged in dying a thousand deaths while reading about the plight of the disadvantaged that one can scarcely collect oneself to take out one's checkbook. Alternately, one may treat so offhandedly the possibility of flubbing a line in one's paper that one neglects even to review the arguments therein, much less rehearsing one's delivery of them. In each of these cases, one's level of involvement with the imaginative object is either too deep or too superficial relative to other, more pressing considerations.

What considerations? The first example, in which one fails to register the import of another person's serious crisis, suggests the violation of a moral norm of conduct, that one should be responsive rather than insensitive to another's suffering. But in the second through fourth examples, some different requirement of proportion seems to have been violated. For instance, responsiveness to another's suffering that is so excessive that it incapacitates one from acting does not seem to exhibit any of the familiar *moral* defects of character. We pity a person who has a nervous breakdown in response to the political torture of his countrymen; we do not condemn him.

What all of these examples have in common is instead the violation of certain *psychological* norms. In each of them, the symmetrical balance between preserving the unity and rational integrity of the self against external violation on the one hand, and maintaining a self-enhancing connection and receptivity to external input on the other, has been destroyed. I discuss this problem in greater depth in Chapters VIII and IX below. In each example, the involvement of the self in its imaginative object is inappropriate because it fails to recognize and respect the ontological boundaries either of the self or of the imaginative object. An appropriate level of involvement in an imaginative object recognizes and respects both

(a) the psychological boundaries of one's self as an acting subject;

and

(b) the psychological boundaries of the other's self as an acting subject.

(a) and (b) apply to cases in which one's imaginative object is another subject. They also apply to cases in which it is not, on the assumption that one's level of involvement in the object itself has consequences for other subjects. The application of these criteria can be illustrated by reconsidering the preceding examples in its light.

The first case described above, in which a written description of others' misfortunes scarcely registers in one's consciousness, much less moves one to action, violates (b), for in it one fails to recognize the existence of the other's subjectivity altogether. This brand of self-absorption comes closest to the primitively egocentric and narrowly concrete view of others described in Section 2. In this case, however, the mental representations of others' interiority exist at least as

surface objects of imagination, while one's own are depth objects. One regards other people as mere furniture in the external environment, and is without a visceral comprehension of their internal conscious states. When we lack a visceral comprehension of what we read, the text in question is a conjunction of empty words without personal meaning to us. Our intellectual grasp of the material is impeded by a failure of the modal imagination those words are intended to spark.

By contrast, the second case describe above, in which one cannot sleep for anxiety at the possibility of flubbing a line in one's paper violates (a). Here the mere possibility of an event that is temporally external to the self in its present state invades that self to the point of disrupting its interior equilibrium. That interior equilibrium itself is treated as a surface object of imagination, whereas the envisioned possibility is a depth object. In such cases, one's preoccupation with external events or anticipated external events is so all-encompassing that one fails to notice one's own internal discomfort at all. This is an abdication of the present self to an anticipated future scenario.

The third case, in which one experiences the agony of the unfortunate one is reading about to such an extent that one is rendered incapable of action, also violates (a), for here, a spatiotemporally external event is allowed to invade the self in its present state to the point of disrupting its interior equilibrium. In this case, one appropriates others' experience of suffering into the self and replaces one's own responses with it. Whereas a visceral comprehension of others' suffering may motivate one to act, the appropriation of their experience as a replacement for one's own renders ameliorative action impossible. Couples who have experienced the contagious effects of one partner's bad mood may recognize this phenomenon. Taking action to help a sufferer requires one to make a sharp distinction between one's own inner state and the sufferer's. Otherwise one abdicates one's actual self to the imagined self of the sufferer.

Finally, the fourth case, in which one is oblivious to the consequences for others of one's neglect to prepare for a future contingency of one's own behavior, violates (b), for in it one fails to respect the validity of other people's normal expectations. This case treats one's audience's inner states – their justified expectations of a certain standard of performance, their assumptions and hopes of intellectual dialogue or edification – as surface objects of imagination, whereas one's own inner state – of confusion, oblivion, complacency, presumption, sloth, or self-indulgence – is a depth object. In this sort of case one fails to imagine with sufficient vividness the difference between others' inner states and one's own. Indeed, one identifies others' inner states with one's own. Like the first, this case illustrates a species of self-absorption that approaches the primitively egocentric and narrowly concrete view, described earlier as resulting from a lack or failure of modal imagination.

In general, then, an inappropriate level of imaginative involvement that violates (a) tends to abdicate the actual, present self to the imagined object. Call this a state of *vicarious possession*. One can be vicariously possessed by the thought of an actual or possible external event as well as by that of another person's inner states. (The possession is vicarious rather than actual because

abdication of the self is in part voluntarily effected.) By contrast, an inappropriate level of imaginative involvement that violates (b) tends to express a failure to modally imagine the object as separate from the self altogether. This draws one closer to the primitively egocentric and narrowly concrete perspective earlier described. Call this a state of *self-absorption*.

Vicarious possession and self-absorption are both a matter of degree, and each can take a variety of imaginative objects. I may be so self-absorbed in my experience of your discomfort as I conceive it that I am completely insensitive to your discomfort as you experience it in fact: Obsessed with reassuring you that your recent auto accident is not likely to reoccur, I fail to notice that my repeatedly broaching and dilating upon the topic only increases your anxiety. Conversely, I may be so vicariously possessed by your conception of me as I envision it that I am completely insensitive to the discomfort it actually causes me to conform to it: Inspired to feats of strength by the conception of me as physically powerful I imagine you to have, I pull unnoticed and uncounted muscles lifting the heavy objects of which, so I imagine, you think me capable. In all such cases, one is self-absorbed by one's own inner state if others' have little impact on it, and vicariously possessed by another's inner state if one's own has little impact on it. Someone who is self-absorbed has too little imagination regarding externals, whereas one who is vicariously possessed has too much.

Vicarious possession and self-absorption are also relative to the actual psychological boundaries of the particular self in question. The self is always constituted by (among other things) the particular social and cultural norms instilled in the process of socialization, as well as by the values, goals and practices that distinguish it both as an individual self and as a member of a specific social community. So what counts as vicarious possession or self-absorption for one self might be a healthy expression of another's central interests or commitments. For example, a self unconditionally devoted to the problem of feeding the starving in India would satisfy the above criteria if it were Mother Teresa's, but would violate (a) if it were Faye Wattleton's; a self preoccupied by memories of its own past experiences might satisfy these criteria if it were James Baldwin's, but would violate (b) if it were Richard Nixon's. The boundaries of some selves circumscribe primarily other-directed or self-sacrificial ideals, whereas those of others circumscribe primarily self-directed ones. Perhaps the more numerous and familiar selves – those that cement most human communities – contain both, in proportions varying with their roles and positions in the community as well as their personal aptitudes and inclinations.

We must first know these facts about their individual commitments and relations to the surrounding community, in order to ascertain whether any particular self is vicariously possessed, or self-absorbed, or both. Cases in which valuable contributions to the world are offset by neglect of loved ones at home furnish numerous illustrations of selves unbalanced by self-absorption in some areas and vicarious possession in others. Take Paul Gauguin, who abandoned his family to go off to the South Seas to paint. His psychological profile gives clear evidence of self-absorption, both in his neglect of his family and in the patent racism and sexism of his attitudes towards the subjects of his painting. On the other hand, his obsession with the

island culture of Tahiti and of his own role in it might be viewed as evidence of vicarious possession, in his abdication to it of the self formed by his prior, longstanding social and familial commitments. Merely his central and overriding commitment to his art by itself – independently of the psychological and social attractions of his adopted as compared to his original environment – cannot be cited as evidence of one or the other, since such a commitment might have existed independently of or concurrently with both.

> There are other such cases, such as Dickens' Mrs. Jellyby in *Bleak House*:
>
> Mrs. Jellyby ... devotes herself entirely to the public. She has devoted herself to an extensive variety of public subjects at various times and is at present (until something else attracts her) devoted to the subject of Africa ... Mr. Jellyby ... is ... merged – in the more shining qualities of his wife. ... [Her eyes] had a curious habit of seeming to look a long way off. As if ... they could see nothing nearer than Africa! (Chapter IV).

It appears that Mrs. Jellyby is self-absorbed, in that she is unable to imagine proximate others (children, husbands, friends) as selves separate from herself; and vicariously possessed by the numerous and transient causes to which she devotes all her energies.

4. Compassion

Next I argue that when the imaginative object is another's suffering, a compassionate response is the symmetric mean between these two extremes.

4.1. Empathy

An involvement with another person's inner states as an imaginative object requires more than that one verbally ascribe certain drives, feelings and thoughts in order to explain her behavior. To do only this much would be to treat those states as a surface object, and so violate 3(b). In addition, it requires that one empathically experience those drives, feelings and thoughts as one observes her behavior. To *empathize* with another is to viscerally comprehend the inner state that motivates the other's overt behavior, by experiencing concurrently with that behavior a correspondingly similar inner state oneself, as a direct and immediate quality of one's own condition. Empathy, in turn, requires an imaginative involvement with the other's inner state because we must modally imagine to ourselves what that state must be as we observe her overt behavior, in order to experience it in ourselves.

These inner states are not to be identified with those one experiences in reaction to her behavior – for instance, as I experience gratitude in reaction to my interpretation of your action as beneficent. Instead they are the inner states that constitute one's interpretation of her behavior – for instance, as I empathically experience subliminal sensations of pain in interpreting your wincing, grimacing, and putting your hand to your forehead. The claim is that an involvement with another person's interiority as an imaginative object is mutually interconnected with one's ability to experience empathically an inner state similar to that which one ascribes to the other as an interpretation of her behavior.

That understanding another person's inner state requires one's empathic experience of it may seem to be a very strong epistemic claim. It implies that understanding another person's inner state – as opposed to merely explaining it – is dependent on a felt psychological connection with the other in a way that understanding a nonpsychological course of events or state of affairs is not. This claim is not as radical as it may seem at first. In Section 1 I argued that modal imagination of another person's inner states as a way of understanding the other person is the norm in most human interactions, without which they all would have a very different cast. In this section it transpires that modal imagination requires, not merely that we envision the other's inner state in order to understand it, but that we viscerally comprehend what we envision as well.

This is no cause for alarm. The implications that there then must be much about other people that transcends our relatively parochial powers of understanding; that we then must work quite hard in order to achieve that understanding, of anyone; and that many human interactions are corrupted by a failure of that understanding should not be surprising and should not be news. I discuss the consequences of moral corruption and the failure of motivational understanding at greater length in Chapter IX below.

How similar one's own state or condition must be to the other's actual inner state, in order to count as a case of empathy, depends on the proportional relations between the intensity and quality of (i) the other's self and his condition, and (ii) one's own self-conception (as defined in Chapter I.7.1) and one's own condition. If you are being disemboweled by a charging bull, and I experience in response only the mildest twinge in my gut, I probably am not empathizing with your condition. Similarly if you are mildly apprehensive about your first driving lesson whereas I am beside myself with panic. These responses of mine fail to count as empathic because they are too different from your actual inner state to enable me validly to attribute them to you.

The more radically I get it wrong when imagining the analogue of your inner state in myself, the less I succeed in understanding yours. The less I succeed in understanding yours, the more the coordination of our actions must depend on convention or force or detailed verbal agreement. And the more we must depend on these factors to coordinate our actions, the more closely we will approximate a "dead-end" relationship of the kind earlier described. Empathy requires, not only a rich modal imagination, but also an approximately accurate one.

How does one achieve empathy without having had first-personal direct experience of that state one attempts to approximate imaginatively oneself? We can only speculate on the extent to which some such external perceptual cues, such as the sight of another person laughing with joy or grimacing in pain, or the sound of a baby crying, might function as biologically ingrained stimuli to which we are biologically disposed to respond empathically. Or we may see another behave in a certain way often enough, and in a sufficiently wide variety of circumstances, that we develop an empathic appreciation of her motives through inference, analogy, or induction. Psychopaths are characterized by, among other things, the inability to respond in

these ways; and we do not yet know whether their disability is primarily social or biological in origin.

However, it is at least clear that forms of creative expression such as music, art, poetry, fiction, and first-person narrative accounts enhance our ability to imagine modally another's inner states, even if we have had no such first-personal experience ourselves. In bypassing mundane third-personal observation of behavior and appealing directly to our interiority, fresh combinations of images, words, metaphors, and tonal progressions enable us to construct an imaginative vision that may in turn causally transform or enlarge our range of emotional responses. Claims that one cannot understand, for example, what it is like for a woman to be raped if one is a man, or what it is like for an African American to be the object of racial harassment if one is European American, have the virtue of refusing to appropriate the singularity of another's experience into one's necessarily limited conception of it. But they are too often based on a simple lack of interest in finding out what it is like, through exploring the wide variety of literary and artistic products designed precisely to instruct us about these things. These creative products may instruct one about another's inner states by depicting what it *would* be like *for oneself* to have those states; or, alternately, what it *would* be like *if one were the other* and had them. But they aid in the cultivation of one's capacity for empathy to the extent that they ultimately enable one to understand viscerally what it *is* like *for the other* to have them. That is, they satisfy both 3(a) and 3(b), above. It is not surprising to find a failure of modal imagination of another's inner states accompanied by a failure of curiosity about them, nor to find an egocentric and narrowly concrete view of others accompanied by a lack of interest in the arts.

We can confirm (to varying degrees) whether or not a person genuinely empathizes with another only by looking at the behavior that inner state is presumed to motivate. But words and deeds alone constitute neither a necessary nor a sufficient requirement of empathy itself, since they might mask the clever dissembler, manipulator, or psychopath. There is no necessary link between the behavior taken as evidence of empathy and the inner state that is empathy. Then how can we know how accurate our empathic responses are? We cannot, since – as I argued in Volume I, Chapter IV.1 – we have no way of comparing interpersonally our own first-personal experiences – even our first-personal experiences of another's inner state as we modally imagine it – with the other's inner state itself. *A fortiori*, we have no way of comparing interpersonally two such first-personal states with respect to quality or quantity.

Nevertheless, we may make rough and ready estimates of the accuracy of our empathic response, by gauging the other's reaction to those of our own actions motivated by it. We may be motivated to respond verbally or behaviorally in such a way that the other's response to our words or actions tells us whether or not they expressed genuine insight into his inner state as we empathically imagined it. Or we may simply ask whether the conjunction of words, phrases, similes, metaphors and colloquial expressions we used in order to describe it is, in fact, accurate; and correct our description and so our understanding according to the other's response. The deep philosophical problems of private language, other minds, and solipsism do not necessarily

engender correspondingly deep practical problems when the effort to understand another is committed, persistent and sincere.

And of course that we cannot know with certainty how accurate our empathic responses are does not imply that there is no fact of the matter about this; nor, therefore, that we cannot approximate empathic accuracy to varying degrees whether we know with certainty that we are doing so or not. We achieve veridical empathy – as well as foresight, clairvoyance, and what is often misdescribed as "extra-sensory perception" – through a capacity of mind that, in Kant's taxonomy, resists the very possibility of independent systematic research, namely intuition. For Kant, *intuition* is the capacity by which we are brought into unmediated relation with an object. This unmediated relation is a precondition of our organizing it in space and time, and of our recognizing it conceptually – and so a precondition of our interpretation of the object as independent of ourselves. But since intuition of another's inner state does not constitute knowledge of it, we may experience what is in fact a veridically empathic response to another's inner state without being able to know, in the strict sense, that we do. I say more about Kant's concept of intuition elsewhere. For present purposes in what follows, I shall often speak of an (accurate) empathic understanding of or insight into another's inner state, as though such a thing is possible. This reflects my Kantian conviction that veridical empathy is not only possible but often actual, even if we cannot know that it is, or how it is.

4.2. Sympathy and Empathy

By contrast with empathy, to *sympathize* with another is to be affected by one's visceral comprehension of the other's inner state with a similar or corresponding state of one's own, and also to take a pro-attitude toward both if the state is positive, and a con-attitude towards them if it is negative. In order to feel sympathy for another's condition, one must first viscerally comprehend what that condition is. Therefore, sympathy presupposes at least a partial capacity for empathy. But once one has achieved an empathic interpretation of the other's behavior, sympathy is of course not the only possible response. I may interpret your behavior as murderous rage, or incestuous lust with the help of my empathic experience of it, and react with even greater revulsion against it for that reason. Whereas sympathy implies one's emotional accord with the other's inner state, empathy implies only one's visceral comprehension of it. That an interpretation of another's inner state requires an empathic imaginative involvement with it does not mean it requires one's concordant reaction to it as well.

An empathic imaginative involvement with another's inner states treats those states as depth rather than surface objects of imagination. It is an application of modal imagination to a particular kind of imaginative object, namely a human subject; and to a particular quality of that kind of object, namely her inner states. To entertain another's inner state as a surface object of imagination is also an exercise of modal imagination, therefore might suffice for mere verbal ascription of inner states to explain another's behavior. But it is insufficient for empathic

understanding of that behavior. An involvement with another's inner states as an imaginative object requires that one empathically experience those states as well.

An inappropriate involvement that violates 3(a), i.e. vicarious possession, has this feature to an excessive degree. In the case of vicarious possession by another person's inner states, one treats one's own inner states as surface objects and the other's inner states as depth objects. To appropriate the other's experience as one imagines it into one's self and replace one's own with it is to

> (1) empathically experience the other's feelings as one imagines them to the exclusion of one's own reactions to them (i.e. a case of being "out of touch with one's feelings");
>
> (2) be so preoccupied with imagining what the other is thinking that one's own thoughts are temporarily suppressed;
>
> (3) act in a way that reflects one's conception of the other's wishes or desires as to how one should act or what should be done.

In general, to be vicariously possessed by another person's inner states means that one's own sentience, rationality, and agency are suppressed in favor of the other's as one empathically imagines them to be. This constitutes an abdication of one's self to another as one imagines him.

By contrast, an inappropriate involvement that violates 3(b), i.e. self-absorption, lacks this feature entirely. When another's inner states are treated as surface objects in deference to one's own as depth objects of imagination, the constituents of one's interpretation of her behavior are empty words at best (assuming one bothers to interpret her behavior at all). Terms such as "headache", "grief", or "starvation" fail to elicit in one any corresponding empathic response altogether. This is one state of mind that makes it easy to toss the letter from the charitable concern into the trash. The moral term for this condition is "callousness", and it constitutes a sacrifice of another's inner states as one conceives them to one's absorption in one's own.

4.3. Symmetry

The contrast between both of these brands of inappropriate imaginative involvement and an appropriate one is that in the latter case, one manages to retain the empathic experience of the other's inner state and the reactions that constitute one's own simultaneously and with equal vividness, in such a way that neither 3(a) nor 3(b) is violated. One holds two equally vivid and sharply distinct experiences – one's own response and the other's as one empathically imagines it – in mind simultaneously. An appropriate imaginative involvement in another's inner state is *symmetrical* with respect to the relation between that state and one's own.

Now it might seem that in so far as this is possible, it would engender agent paralysis. It might seem that to empathically imagine to oneself another's inner state with a vividness equal to one's direct experience of one's own would be to be torn between being motivated to act by the other's inner state as one empathically imagines it, and being motivated by one's own inner state as one directly experiences it. If I empathically imagine you to experience embarrassment at the

same time and with the same vividness as I directly experience *Schadenfreude* in response, then it appears that neither motivational state overrides the other in my consciousness. Then what spurs me to act at all?

However, this difficulty is more imagined than real. First, these two states may be equally vivid without being equally intense. The *vividness* of an object or state depends on its perceptual (not necessarily visual) clarity, and on the sharpness of its sensory detail. The *intensity* of a state depends rather on the strength of its causal impact on one. For instance, your heady pride of achievement may meet with only faint enthusiasm in me. Yet I may empathically imagine your heady pride of achievement no less vividly than I directly experience my own faint enthusiasm for it. Second, that I experience simultaneously and with equal vividness two different motivational states does not imply any further similarity of structure between them. A structural feature that my own inner state has and that my empathic imagination of yours lacks is a direct connection to my own capacity for agency. Whereas I can empathically imagine your inner state, I cannot spur you to action on the basis of my imaginative involvement with it. By contrast, my direct experience of my own inner state in response *can* spur *me* to action on the basis of my imaginative involvement with *it*. Essential to the boundaries that enable me to distinguish my self from yours, hence to satisfy 3(a) and 3(b), is the natural link between my self and my action that is missing between your self and my action, or between my self and your action.

It is only when this natural link is weakened that violations of 3(a) or 3(b) occur. For example, when a child is repeatedly told that he feels what his caretakers think he should feel instead of what he does feel, he may learn to suppress awareness of his own responses and replace them in imagination with others that are prescribed to him. This habit of thought encourages vicarious possession. Alternately, when others regularly assume responsibility for a child's actions and shield her from their human consequences, she may fail fully to develop the capacity to imagine modally others' responses to them as independent of her own wishful thinking about them. This habit of thought encourages self-absorption. Both of these cases involve a conflation of one's own interiority with that of others, and so a severance of the natural link between one's own thought and one's actions. In the first case, of vicarious possession, one's own action is guided by another's conception, as one empathically imagines it, of one's own inner state. Such a case can lead to agent paralysis when I empathically imagine your conception of my inner state to be at least as motivationally compelling as my direct experience of my own response in fuelling my action. In the second case, of self-absorption, one's action is guided by one's own conception of another's inner state as one self-centeredly imagines it. Such a case can lead to agent paralysis when I imagine my conception of your inner state to be at least as motivationally compelling as your direct experience of your own in fuelling your action. In neither case, however, do I succeed in directly experiencing my own inner state as fuelling my action with the same vividness and intensity as I empathically imagine your inner state as fuelling yours. Only in this last case is neither 3(a) nor 3(b) violated.

Of course my empathic imagination of your inner state as comprising a desire that I act in a certain way can spur me to action, but only if I already directly desire to act as you desire me to act. Or, my empathic imagination of your inner state as comprising a desire to act in a certain way can spur me to action, but only if I mistakenly imagine, empathically, that I am you. But both of these possibilities violate 3(a). The first abdicates my self to the desire, which I empathically imagine you to have, that I act; my original desire to act as you desire me to act is ignored. The second abdicates my self to the self I empathically imagine you to have. Both possibilities require a severance of the direct connection between my capacity for agency and my own inner motivational state. Both possibilities require establishing a connection between my capacity for agency and the motivational state I empathically imagine you to have. Thus both require my vicarious possession by your inner state as I empathically imagine it. This just is to appropriate your responses into my self and replace it with them. It is to treat my own inner state as a surface object of imagination, and your inner state as a depth object. It is not to treat both as occurring simultaneously and with equal vividness after all.

Alternately, my primitively egocentric conception of your inner state as comprising a desire that you act in a certain way can spur you to action, but only if you already desire to act as I imagine you desire to act. Or, my primitively egocentric conception of your inner state as comprising a desire to act in a certain way can spur you to action, but only if you mistakenly imagine, empathically, that you are me. But both of these possibilities presuppose a brand of self-absorption on my part that violates 3(b). The first sacrifices your self to the desire to act that I egocentrically conceive you to have. The second sacrifices your self to the self you empathically imagine that I conceive you to have. Both possibilities require a severance of the direct connection between your capacity for agency and your own inner states. Both possibilities require establishing a connection between your capacity for agency and the inner states you empathically imagine me egocentrically to conceive you as having. Thus both require your voluntary submergence in my imaginative but primitively egocentric reconstruction of your inner state. This imaginative reconstruction treats my own inner states – including those I egocentrically conceive you to have – as depth objects, and your actual inner states as surface objects of imagination. Again the symmetry required of an appropriate imaginative involvement is lost.

When the other's experience is one of suffering, the appropriate imaginative involvement that satisfies both 3(a) and 3(b) is one of *compassion*. Compassion comprises at least three further distinguishable responses. First, it includes empathic understanding of the other's condition. Second, it includes sympathetic "fellow feeling" in reaction. And third, it includes a consequent disposition to render aid or show mercy to the other. So compassion includes cognitive, affective, and conative components respectively.

To render aid, mercy or restitution to another is not the same as acting unreflectively on a momentary feeling of concern. It is rather to act consistently and reliably in such a way calculated to relieve the other's distress. That is, it is to act in accordance with a normative

principle of moral conduct that itself has application to a variety of situations. By contrast with occasional stirrings of sympathy that may or may not spark fleeting impulses to help, compassion is a principled, transpersonal moral emotion that moves one to a course of action in accordance with a normative requirement of rendering aid. As is the case with all normative moral principles of conduct, the requirement to render aid is a requirement that one strike a symmetrically balanced accommodation between the condition and demands of the self and the condition and demands of another.

Striking a symmetrically balanced accommodation between these two different sets of interests and demands requires that the self be vicariously possessed by neither, but that it have a deep imaginative involvement – one that is antithetical to self-absorption – with both. Vicarious possession by the other's inner state would constitute a sacrifice of the integrity of the self to the inner deprivation or suffering of the other. It would be to take on the other's suffering as an internal condition of one's own. This would mean paralyzing or incapacitating oneself, in the ways earlier described, from consistent and principled agency in the service of relieving that suffering. When altruistically inclined agents worry that an active, participatory commitment to solving an intractable social problem (such as inner city poverty) will "suck them dry," it is the fear of this very real kind of incapacitating self-sacrifice that they express. But incapacitating self-sacrifice, and the sacrifice of one's own needs and interests that accompany it, is a consequence of vicarious possession by the other's suffering. It is not a consequence of compassion properly understood.

As defined in this discussion, compassion precludes such abnegation of the self and its interests because compassion disposes one to act in accordance with the moral principle of rendering aid to the needy. Applying this principle requires one to conceive of oneself either as a potential provider or as a potential recipient of aid, and calls upon the former to put their resources in the service of the latter. But incapacitating self-sacrifice is clearly a condition of need that itself demands amelioration. Hence consistent application of the principle of rendering aid to the needy prohibits depleting or sacrificing one's resources so thoroughly that one ends up joining the ranks of the needy oneself. Rather, the terms of this principle implicitly require protecting the psychological integrity of the self that is disposed to act on it, at the same that it requires extending the self in service to the other. So the principle of rendering aid to the needy imposes a double requirement of balance on the affective and conative dispositions it regulates.

Compassion satisfies the double requirement of balance by satisfying the symmetry requirement already discussed. Indeed, this double requirement just is a special case of the symmetry requirement. In compassion, the interests and demands of the self are balanced in relation to those of the other because the self as a unified whole is balanced in relation to the other. The self is situated between self-absorption and vicarious possession with respect to another's inner state of suffering. It is a condition both of inviolate inner integrity and of experiencing the other's felt distress, in which the demand for relief of that distress is met by principled action to restore the other to a condition of similarly inviolate integrity.

Mean-spiritedness, by contrast, evinces poverty of spirit. It is a condition of emotional deprivation in which inner integrity is violated by the other's felt distress – i.e. in which one is vicariously possessed by that distress; and in which the demand for relief of that distress is met by desensitizing and fortifying the self against it – i.e. in which one is self-absorbed by one's own. Thus the spiritually undernourished or mean-spirited self swings between vicarious possession and self-absorption relative to the other's distress. It is bereft of the inner resources both for preserving the integrity of the self against incursion by the other, and for extending those resources beyond the self to the other.

Whereas compassion presupposes the integrity and emotional abundance necessary to fuel actions on behalf of another as well as those on behalf of oneself, mean-spiritedness involves a felt violation, an emotional deficit in which action of behalf of the other is experienced as an extortion, as usurping those on behalf of oneself. Compassion thus prepares the self for a balanced accommodation with the other because it requires one neither to sacrifice one's own well-being on the other's behalf, nor the other's well-being on one's own. Instead it involves respect for the psychological boundaries of both, and a disposition to restore the inner integrity of the other that is altruistic without being – literally – self-sacrificial.

This is why compassion requires a symmetric imaginative involvement with the other's inner states. Unlike both vicarious possession by another's suffering, which violates 3(a), and self-absorption, which violates 3(b), compassion preserves the symmetry, required of an appropriate imaginative involvement with another's inner state, between one's empathic understanding of that state and one's own direct reaction to it. In compassion, I sympathetically feel the same inner state I empathically imagine you to feel, namely suffering, and with the same vividness I imagine you to feel it. However, my sympathetic experience of your suffering as I empathically imagine it is connected to my agency in a way in which your direct experience of your suffering as I empathically imagine it is not. That my sympathetic experience is of *your* suffering as I empathically imagine it, and not of my own, is what inclines me to ameliorate your suffering rather than my own. That my sympathetic experience of your suffering as I empathically imagine it is *sympathetic* is what inclines *me* to ameliorate your suffering rather than (or in addition to) you. And that my sympathetic experience is of your *suffering*, rather than of your gratification, is what inclines me to ameliorate it rather than promote it.

But if my sympathetic experience is overwhelmed by the vividness and depth of your suffering as I empathically imagine it, then I abdicate my sense of self and agency to the self I empathically imagine you to have; I am vicariously possessed by your suffering. And if your suffering as I empathically imagine it is overwhelmed by the vividness and depth of my sympathetic experience of it, then I sacrifice your suffering as I empathically imagine it to my sympathetic experience of it; I am absorbed in that sympathetic inner state of my self I empathically imagine to be yours. Like dead-end relationships, self-absorption in one's own sympathy for others is hardly an unfamiliar phenomenon; but it is itself more worthy of pity than sympathy. That is why an imaginative involvement with another's suffering counts as

compassion only if it is symmetric with respect to the relation between the other's empathically imagined inner state and one's own sympathetic one.[3]

5. Blum's Argument Against Impartiality

Now to take up in greater detail Blum's characterization of impartiality as being unbiased by one's personal preferences or interests in one's treatment of others. Blum adds that it involves "giving equal weight to the interests of all." (44) Presumably he means "equal weight *other things equal*," since, as we saw in Section 1, it would be a sign of bias, not impartiality, to give equal weight to the interests of the homeless and to those of billionaire real estate developers in distributing governmental funding for affordable housing, when the interests of the homeless weigh so much more heavily. We can say, then, to begin, that to be impartial is to treat competing preferences and interests on their own merits and without being biased by one's own. Even with this adjustment, impartiality remains a metaethical requirement rather than a normative moral principle, since we must first know what these interests are and for what they are competing – information provided in the normative principle to be applied – in order to identify the nonarbitrary attributes relative to which the principle can be impartially applied. In all such cases the requirement of impartiality directs us to apply a normative principle of conduct evenhandedly. It does not tell us which normative principle to apply. In these three concluding sections I show that compassion requires not only a symmetric imaginative involvement with another person's interiority, but therefore a disposition to impartiality of treatment as well.

Clearly, impartiality as just characterized presupposes modal imagination. It requires one to imagine as depth objects interests and preferences that one may not have, and may never have had. This requires of one an imaginative involvement with the inner states of those who have them. As we have seen, such an involvement is a necessary condition of the ability to form universal concepts of inner states such as love, fear, desire, or joy – concepts that extend backward into a counterfactually possible past and forward into a possible future. Modal imagination is what enables one to apply these concepts to instances of possible in addition to actual experience, and so to apply them to the imagined inner states of others of which one has no actual experience at all.

Without an empathic imaginative involvement, one's understanding of the interests and preferences of others would remain purely verbal; they would be surface objects of imagination. This is not to maintain that they would be entirely lacking in significance. But one would lack insight into what was at stake psychologically and emotionally for individuals who have those preferences and interests. By contrast, to the extent that one had first-personal insight into what was at stake psychologically and emotionally in having one's own preferences and interests, those interests would be depth objects of imagination. In thus violating symmetry, one's capacity for impartiality would be correspondingly defective. One's judgment would be distorted by the psychologically and emotionally compelling representation of one's own interests and preferences, relative to which others' would appear by definition less compelling.[4] The same

argument applies when we must judge impartially, not between our own interests and another's, but between two third-personal sets of interests, in only one of which we have an imaginative involvement.[5]

We may begin, then, by thinking of impartiality in the judgment of preferences and interests as the result of applying a universal and general, normative moral concept or principle to those relevantly situated agents' inner states selected by the terms of that principle, such that the inner states of the person applying the principle do not lead him to tailor its application to his own situation, nor add special weight to his personal interests or allegiances in determining its application; this just is the conception of impartiality defended in Volume I, Chapter VIII.3.2. So, for example, an impartial application of the principle of directly apportioning quantity of resources to need in the distribution of governmental funding for affordable housing would not give any special weight to the need of the distributor to cement her political alliances. Nor would it tailor the application of this principle to her personal or social connections to billionaire real estate developers. An impartial application of this principle would compare the respective inner states of need of all designated parties relative to one another, on the basis of a symmetric, empathic imaginative involvement with those of each, and distribute the funds accordingly.

Such a distribution presumes no solution to the problem of interpersonal comparisons discussed in Volume I, Chapter IV.1, since a symmetric empathic understanding of another's inner states does not aspire to the objective quantifiability of those states. Indeed, the irreducibly qualitative variety among such states precludes this. As suggested in Section 4, it assumes, without being able to show or prove, the capacity of one's modal imagination to subjectively represent as depth objects the quality and intensity of others' inner states with some degree of *de facto* accuracy. This capacity is based on an empathic comprehension of the behavior that ordinarily accompanies them, and on rough and ready behavioral interactions that then enable one to fine-tune one's empathic insights. It also assumes one's capacity to preserve the distinctive quality and intensity of each such imaginative object with equal vividness, simultaneously in one's consciousness. It assumes, that is, our ability to experience walking and chewing gum at the same time, even when it is oneself who is doing the walking and another who is chewing the gum.[6] And it assumes one's ability to compare such vividly imagined objects with respect to one's subjective representation of their quality and intensity. In a symmetric empathic understanding of another's inner states, the scale of quantitative calibration among these states as imaginative objects is a function of their relative effect on the subject. It is ultimately the quality and relative intensity of one's own experiences that are being compared.

Some philosophers have offered procedural accounts of impartiality. It has been claimed, for example, that impartiality of judgment is what results from putting oneself in the place of the individual whose preferences are being judged[7]; or that it results from discounting one's own interests and desires when making the judgment[8]; or both. The close conceptual connection between all of these accounts of impartiality and the foregoing analysis of compassion deserve emphasis. Both impartiality and compassion require an empathic imaginative

involvement with the other's inner state, and both require a reduction of the pre-eminence in consciousness of one's own inner state, in order to arrive at a judgment that appropriately balances the interests of the self and those of the other. So both impartiality and compassion require an imaginative extension of the self into the domain of the other's interiority, and a corresponding imaginative accommodation of the other's interiority within the domain of the self. It is difficult to see impartiality and compassion as being as mutually exclusive as Blum seems to think.[9]

However, all of these accounts of impartiality are faulty in presupposing the natural pre-eminence in consciousness of one's own inner states over another's as one empathically imagines them. Each assumes, without explicitly stating this, that impartiality consists in applying a corrective to a natural tendency to self-absorption alone – as though vicarious possession were not as much of a vice, and as prevalent a vice, at the opposite extreme. Consequently, taken at face value, these two procedures, alone or in conjunction, exhibit bias toward the other. Both advocate the suppression of the self in the service of vicarious possession by the other. But the symmetry requirement implies that impartiality could not result from either of these procedures considered independently, or from both of them conjoined, for this very reason. If impartiality requires unbiased judgment, then *the judgment in question must be biased neither toward oneself nor toward the other*. Call this *strict impartiality*. An adequate procedural account of strict impartiality – which I do not purport to offer here – must explicitly steer the self clear both of vicarious possession and of self-absorption.

Blum's rejection of impartiality as appropriate and intrinsic to feelings of compassion seems to stem from the view that impartiality is merely a corrective to a predominantly self-interested tendency to make personally biased judgments about the proper weight to be accorded other's interests in the pursuit of one's own. If this is all impartiality is, then of course it will follow, as Blum seems to infer, that a compassionate person whose judgments are not biased by an excess of self-interested concern has no need of impartiality's corrective influence. But this presupposes the Humean conception of the self as motivated by essentially self-interested concerns, to which impartiality is the corrective and compassion the exception. That is, compassion (as well as friendship and altruism) in Blum's account functions as though it were a counterexample to a generally valid empirical generalization about the *de facto* prevalence of self-interested motivation and judgment biased accordingly.

6. Strict Impartiality

In Volume I of this project, I have gone to some lengths to argue that the Humean conception of the self is inadequate as a descriptive model of human motivation; and in the present volume, that other-directed motives such as sympathy and altruism can play a more central role. Does this mean that we may dispense with strict impartiality as a virtue as well? Clearly not. An altruistic person may give unjustifiably short shrift to his own interests in devoting his energies to others. Or a sympathetic person may be uncertain to whom, among the

many claimants on her sympathy, she should direct her sympathetic response. Strict impartiality has a central role in the analysis of compassion, because so many claims on our sympathy regularly confront us, including those of our own interests and preferences, that we are compelled to adjudicate among them. As we have seen in Sections 3 and 4, a healthy compassionate response to others demands that we navigate between the Scylla of self-absorption and the Charybdis of vicarious possession. It demands that we find a principle for distinguishing between unhealthy fortifications or transgressions of the boundaries of the self, and healthy social expressions of it. A principle of strict impartiality meets this demand.

The symmetry requirement on compassion as an appropriate imaginative involvement with another's suffering implies that compassion presupposes strict impartiality of modal imagination. We have already seen in Section 4.3 that unlike occasional and unpredictable stirrings of concern, or impulsive attempts to be helpful, compassion involves a disposition to respond to the suffering of another in a consistent and discriminate manner, i.e. in accordance with universal and general normative principles of aid, mercy, or restitution that, *like all normative moral principles*, require a symmetrically balanced accommodation of the demands and interests of the self with those of the other. Compassion achieves such an accommodation by avoiding both vicarious possession by the other's distress and self-absorption by one's own, and so by disposing the self to action that sacrifices the inner integrity of neither self nor other.

Moreover, satisfaction of the symmetry requirement implies that compassion as a moral motive is consistent with personal dislike or revulsion toward the object of one's compassion, because the empathic comprehension of the other's suffering, the sympathetic reaction to it, and the respect in which compassion disposes one to extend oneself on the other's behalf in order to ameliorate it, is independent of attributes irrelevant to those picked out by the principle of rendering aid to the needy. Where personal dislike of the sufferer precludes sympathy with his distress, symmetry is violated, skewing the self toward self-absorption; and bias thereby precludes compassion from taking hold. Impatience with the other's personal vanity or disgust at his malodorous garb may coexist with the feeling of compassion because the object of that feeling is his suffering and his need, not his self-estimation or his sartorial habits; and because the resulting disposition to action is directed to the amelioration of his suffering and his need, not to the improvement of his personality or sense of style. Strictly impartial conformity to an inherently impartial, substantive prescriptive principle of compassion rules out as attributively irrelevant both sacrifice of self or other in the amelioration of suffering, and also bias toward popular or charming sufferers over unpleasant or socially repulsive ones.

The strictly impartial application of such principles thus requires an absence of personal bias, both toward the other's inner state and toward one's own. One exhibits personal bias toward another's inner state to the extent that one's imaginative involvement with it is weighted toward vicarious possession: one appropriates the other's suffering as one empathically imagines it into one's self and replaces one's own with it, as described in 4.2(1)-(3). By contrast, one exhibits personal bias toward one's own inner state to the extent that one's imaginative

involvement with the other's recedes towards self-absorption, with primitive egocentrism and narrow concreteness constituting the extreme.

But why describe these as cases of personal *bias*, rather than of mere imaginative excess and failure respectively? A bias, unlike a merely unbalanced imagination, presupposes a value judgment, i.e. that the object of bias is more worthy of favor or consideration than the alternative. The basis for this judgment is the possession by the object of bias of some specific but irrelevant attribute that the alternative is perceived to lack. In the case of an imaginative involvement with one's own experience or that of another, personal bias occurs when one evaluates either as more worthy of favor or consideration than the other on the basis of a specific but irrelevant attribute that the one has and the other is perceived to lack. For example, one may regard another's pain as one empathically imagines it as more worthy of consideration than one's own as one directly experiences it, because one regards other people in general as more important or worthy than oneself; or because one regards other people's inner states as intrinsically more interesting or worthy of investigation than one's own. In either of these cases, the irrelevant attribute that directs one's personal bias to the other is the attribute of being other than oneself.

Conversely, one may regard one's own pain as more worthy of consideration simply because it is one's own, or because one regards oneself as in general more important or interesting than others. Unlike cases in which one regards one's own or another's pain as more worthy of favor or consideration *because the pain in question is more intense*, these cases exhibit personal bias because the attributive basis for ascribing superior value to the one or the other is arbitrary and irrelevant. The mere fact that my headache is mine does not entitle it to precedence in my imagination over your imminent demise from malnutrition. Nor does the mere fact that your suffering is yours entitle it to precedence in my imagination over my sympathetic response to it. Indeed, if my sympathetic response to your suffering is to motivate my ameliorative action on your behalf, your suffering as I empathically imagine it had better *not* overwhelm my sympathetic response to it.

Of course it might happen that the pain of my sympathetic response to your suffering is greater than the pain of your suffering as I empathically imagine it. Conceiving of myself as infinitely more sensitive than thou, I might suffer for you in a way that I empathically imagine you to be incapable of suffering yourself. Hence this is a case not of vicarious possession but rather of *surrogate martyrdom*. Surrogate martyrdom is distinct from genuine martyrdom because a genuine martyr shoulders the actual suffering of others, not the suffering she imagines they would feel were they as sensitive as she. Since greater pain justifies greater consideration, according to the foregoing account, surrogate martyrdom would seem to warrant more attention to my sympathetic response than to your suffering, without implying personal bias. However, in conceiving of myself as being more sensitive to suffering than thou, I violate 3(b), for I imagine your inner state of suffering as though it were a surface object of imagination in comparison to my own inner, sympathetic state as a depth object. Hence even surrogate martyrdom implies personal bias. The bias consists in arbitrarily ascribing superior sensitivity to myself and

weighting my imaginative involvement accordingly. Surrogate martyrdom is therefore distinct from genuine compassion.

What about the standard case, in which the magnitude of your pain as I empathically imagine it exceeds the magnitude of my sympathetic response to it? Since neither 3(a) nor 3(b) is violated, surely symmetry is violated by our unequal experiences of pain, without implying personal bias in this case? Not so. This standard case is analogous to that discussed in Section 4, in which your heady pride of achievement outstrips my faintly enthusiastic response to it, and the answer is the same. I may hold in mind with equal vividness both your greater pain as I empathically imagine it and my lesser sympathetic pain-response to it. Symmetry remains inviolate, and therefore strict impartiality does as well. Compassion has the psychological feature that neither the other's suffering as one empathically imagines it nor one's own sympathetic response to it is submerged by the other, regardless of the magnitude of either.

This analysis extends to third-person cases. Consider, for example, the friendship case Blum raises for discussion. Blum thinks it is obvious that when choosing between helping a friend and helping a stranger,

> (1) one is morally permitted to choose to help the friend simply because he is one's friend.

However, this view has bite only if the stranger is stipulated to be in greater need of help. In that case, as it turns out, Blum acknowledges the possibility that

> (2) if the stranger is in greater need of help, he may have a superior claim on one's compassion. (49)

In these passages, Blum's discussion treats the psychological fact of compassion as generating normative moral principles, among them that the object of this emotion should be the recipient of one's ameliorative action. But the plausibility of this normative principle depends on rejecting the connections between strict impartiality and compassion for which I have argued here. Specifically, Blum's notion of compassion is consistent with the primitively egocentric view of others described in Section 2, according to which one's treatment toward others is determined by how fully they happen to engage one's feelings.

By contrast, my conceptual analysis of compassion, as including satisfaction of the metaethical requirement of strict impartiality, carries no such normative implication. My analysis leaves open the questions whether compassion should be motivationally central in a normative moral theory; whether or not one should act on those principles of aid in a particular case; if so, whether one is most appropriately motivated by feelings of compassion, ties of personal loyalty, or the voice of conscience; and to whom, among the deserving candidates, one should direct one's ameliorative efforts.

Nevertheless, the foregoing analysis can accommodate both (1) and (2) above. Friendship, too, is governed by normative moral principles of conduct and emotion. As in the case of compassion, adherence to these principles requires an empathic imaginative involvement with the other's interiority that violates neither 3(a) nor 3(b). Without satisfaction of these two

conditions, one's relation to the other is poisoned either by vicarious possession or by self-absorption. Vicarious possession by another's inner states bespeaks a level of psychological dependency on the other that is patently inimical to genuine friendship. Self-absorption in one's own inner states or self-serving conceptions of the other bespeaks an insensitivity to and disrespect for the other that is equally antithetical to genuine friendship. So genuine friendship presupposes strictly impartial satisfaction of inherently impartial, normative principles of mutual sensitivity, respect, and psychological independence; and therefore satisfaction of the symmetry requirement. Therefore friendship presupposes strict impartiality. And when a friend suffers, this strict impartiality is expressed in compassion for her condition.

When a friend and a stranger suffer with equal intensity and one empathically imagines the inner states of both with equal vividness, a compassionate person will feel equal sympathy for both, and equally moved to ameliorate the suffering of both. Because the inner state of each bears the same relation to one's own, namely satisfaction of the symmetry requirement, compassion evinces a strictly impartial concern for the stranger's as well as the friend's condition. What finally determines one to render aid to one's friend instead of the stranger is not one's heightened compassion for the friend. What moves one to help the friend are the bonds of mutual trust, loyalty, shared history, responsibility and respect that uniquely define the relation of friendship.

This conclusion departs from Blum's in two respects. First, Blum seems to think that there is a psychological connection between liking someone more, or having a more intimate relationship with him, and feeling greater compassion for him. In Section 4 I rejected this connection, on the grounds that compassion is strictly impartial with respect to irrelevant attributes that might bias one either towards or against the sufferer. But moreover, the psychological connection may work in the opposite way: it may happen that the more intimately one knows a person, the more one becomes accustomed to his suffering, and the more emotionally inured one becomes to it. Hence friendship may undermine compassion rather than promote it.

Second, Blum believes there is a prescriptive connection between having a more committed or intimate relationship with someone and feeling greater compassion for her suffering. I reject this connection on the grounds that it prescribes stronger feelings of empathy and sympathy, and a more motivationally effective disposition to render aid on grounds irrelevant to the magnitude of the pain felt by the sufferer, and irrelevant to the magnitude of her need for aid. That is, it prescribes feeling more compassion for people we know than for people who are in greater pain. I find this prescription unacceptable, but not only because it expresses clear bias towards an attributive basis that is irrelevant for feeling compassion. It is also unacceptably exclusionary in the presence of those for whom the conditions of survival make friendship an unattainable luxury and whose magnitude of suffering clearly surpasses that which anyone we know is likely to experience first-hand. Compassion demands a generosity of spirit that is incompatible with narrow and arbitrary restrictions of scope. So I insist on satisfaction of

the symmetry requirement in compassion for prescriptive as well as psychological and conceptual reasons.

Compassionate action toward *one* other requires only the special link between my self and my action when the symmetry observed is between my own and the other's inner state as I empathically imagine it. By contrast, compassionate action when symmetry is observed between my own and *many* others' inner states also requires, when all suffer equally, some further motivating attribute of the particular other on whose behalf I compassionately act. Since one's own strict impartiality among equally suffering others expresses an inherently *ceteris paribus est* relation among agents, one's compassionate action on behalf of any requires some sort of motivational tie-breaker among them. Otherwise agent paralysis really does set in.

In the case in which the stranger patently suffers more intensely, the dictate of compassion is equally clear: My empathic imaginative involvement with the plight of brutalized African women will move me to contribute funds to Transafrica, rather than to my friend's purchase of a new coat, when these two options conflict, because I perceive the greater intensity of suffering in the former. But the responses to each of these cases are applications of the strict impartiality requirement, not precluded by it. In the first case, strict impartiality determines the empathic recognition of equal suffering on the part of both friend and stranger, and of the bonds and obligations of friendship as a tie-breaker. In the second case, strict impartiality determines the empathic recognition of greater suffering on the part of the stranger despite those bonds and obligations that might otherwise have biased one toward the friend. In both cases, the requirement of strict impartiality fixes one's compassionate response to the situation in such a way as to give one's own interests and attachments no more and no less than their due. Thus the fact that strict impartiality as a metaethical requirement of adequacy on the application of any normative moral principle (not itself such a principle) implies that the fact that one's experience of identifiable compassion for one or many sufferers will *move* one to ameliorate their suffering does not by itself prescriptively commit one to ameliorative action on their behalf: Feelings of compassion may need to be balanced against considerations of efficiency, rational prudence, or other moral obligations – such as those to friends or family, and may not always override them.

The unbiased application of distributive principles, the emotion of compassion, and the relation of friendship are not the only moral virtues that presuppose strict impartiality between self and other. Honesty, trust, love, and responsibility – indeed, any virtue susceptible to analysis in terms of normative principles of behavior – could be treated similarly, although I do not attempt this here. The general point is that strict impartiality requires the ability to balance the demands and interests of the self with those of others in accordance with a normative principle biased toward neither. Indeed, the set of moral principles that constitute a normative moral theory just is a strictly impartial solution to the problems created by the competing demands and interests of different selves; i.e. it solves a Prisoner's Dilemma-type situation.[10] So it is not surprising that Kantians insist that this ability is definitive of the moral point of view, and that it enters into the conception and practical application of every moral virtue. Without strict

impartiality, personal interactions would consist solely in manipulative self-absorption or dependent vicarious possession. Feelings of injustice, violation, neglect or betrayal are moral reactions that rightly alert us to the operation of these vices in our social relationships.

That the functioning of moral virtues such as compassion or friendship presupposes empathic modal imagination of another's suffering which is strictly impartial with respect to the relation between one's own inner state and others' explains why commitment to an impartial moral theory engenders rather than precludes such virtues. I have argued in Chapter V.5.2 that a moral theory is an ideal descriptive theory that enables us to make sense of our moral experience: to identify another's condition as one of suffering, for example; or our own behavior as that of rendering aid. I also argued in Volume I, Chapter VIII.3.2.2., as well as above in Section 1, that if it is a genuine theory, a moral theory is by definition (strictly) impartial, since it contains neither definite descriptions nor arbitrary attributive bias. In this discussion we see how a strictly impartial moral theory might function both to constitute and to regulate our empathic imaginative responses to another's condition in a morally appropriate way.

Moral theory constitutes our imaginative responses by providing us with concepts of morally virtuous – i.e. strictly impartial – character. We use these concepts to identify, understand and evaluate our experiences of our own inner states, as well as those of others' as we modally imagine them. Moral theory also regulates our imaginative responses, in that these strictly impartial concepts of virtuous character serve to guide their cultivation. By describing ideals of character and action against which we compare our own, the strictly impartial concepts of normative moral theory provide criteria of self-evaluation the application of which itself contributes to our moral growth. In applying these criteria we come to understand the difference between, for example, a balanced, sensitive response to another's suffering, versus one that uses another's suffering to meet various unmet psychological needs of one's own. We thereby come to see that what distinguishes compassion from vicarious possession and self-absorption is not the agent's good will toward the sufferer, and not his desire to minimize unhappiness as completely as possible. A person whose responses to another's suffering fail to satisfy the strict impartiality requirement of compassion is not necessarily an immoral person. But we rightly say of such a person that he is infantile, self-indulgent, or lacks vision; or, alternately, that he is too invasive, self-abnegating, or meddlesome to behave reliably as a moral agent. We come to see that what distinguishes compassion from vicarious possession and self-absorption is the more general requirement of a strictly impartial moral theory, that we treat another's moral personhood with no more or less than the care and respect we accord our own – i.e. with the care and respect due a moral person impartially considered.

7. Moral Motivation and Moral Alienation Revisited

But strictly impartial moral theory regulates our imaginative responses in a second respect, by providing the impartial principles that motivate and guide moral conduct. We have seen in Volume I, Chapters VI and VIII that on the Humean conception of the self, any account of

moral motivation to act on impartial principles must either presuppose a desire to act on those principles, in which case my compassionate response to another's suffering is "morally alienated," or else it can issue only from the impersonal point of view of those principles themselves. But on the Kantian conception of the self defended in this volume, the self just is that coherent psychological entity which is constituted and rationally structured by the concepts and principles – i.e. the functions (to use Kant's term) by which lower-order concepts and particulars are subsumed under higher-order concepts – that define its perspective at a given moment. In Chapter V above we also have seen how such rational principles can provide both necessary and sufficient conditions of action. Moral principles that satisfy the requirements of horizontal and vertical consistency over time are rationally intelligible to the agent who holds them, and so are a species of rational principle. Hence on this conception, moral conduct in the ideal case is motivated directly and without mediation by those rational, strictly impartial, specifically moral principles that are partly and necessarily constitutive of the agent's own point of view.

7.1. Motive versus Purpose

Applying these conclusions to the case of specifically morally motivated conduct, reconsider Blum's question as to our moral obligations toward a friend in need. Suppose the case to be that *ceteris paribus est* situation in which the bonds and obligations of friendship are the tie-breaker, such that my symmetrical and strict impartial moral theory condones my rescuing my best friend Ellsworth from drowning first, and before the stranger nearest to me on the sinking ship that holds us all. Let us heighten the Kantian cast of this example by describing it as a case of my being motivated to rescue Ellsworth first, by *respect* for a strictly impartial moral *imperative*, derived from that part of my moral theory which assigns me special obligations to friends, to aid friends first, other things equal, when rendering aid to the imperiled.

I couch this example in terms of respect for a moral imperative in order to show that this analysis can refute the Humean and Anti-Rationalist, even on the most exoteric and commonplace interpretation of Kant's moral psychology. In fact, as indicated in Chapter I, I do not think this interpretation does justice to Kant's concept of *Achtung*; nor, therefore, that talk of respect or, for that matter, of imperatives is fully adequate to the spirit of Kant's view. In addition to the scholarly issue of correct translation, I also object to Beck's and Paton's translations of *Achtung* on aesthetic and strategic grounds: It makes Kant's moral psychology look much more cumbersome and eccentric than it is in fact. I explained in Chapter V how an occurrent thought or belief can directly precipitate rational action, and the same considerations apply here. On Kant's view (and on mine), we in fact do not necessarily experience an emotional response to the normative moral principles that motivate and guide our behavior. Rather, they elicit from us a certain *attitude* toward them – of susceptibility or rapt attentiveness or mindfulness or interest or receptivity. As I suggested in Chapter V.5.1, the moral principles that motivate and guide our behavior compel our attention in the same way that *modus ponens* does. I defend these claims at greater length elsewhere, and have more to say about imperatives in Chapter IX, below. But in

order to address the Humean and Anti-Rationalist criticism on its own terms, I continue to use this exoteric terminology here. Their complaint about this case then would be that I am motivated to rescue Ellsworth first by a desire to obey the imperatives of my moral theory and not by my compassion for Ellsworth.

However, this complaint is mistaken. Recall Kant's distinction between a purpose and a motive for acting.[11] A *purpose* for acting is the goal, end, or intentional object to the achievement of which my behavior is directed. A *motive* for acting is the psychological cause of action, i.e. that which moves me to behave intentionally. Under the Humean influence, most of the philosophers discussed in Volume I assume that the purpose of my action is necessarily its psychological cause as well. They assume this because they suppose that the purpose of my action must be the object of a desire, or, minimally, of a "pro-attitude" toward it, which suffuses it with a weak but rosy glow and inspires me to pursue it. And according to Brandt and Kim's analysis discussed in Volume I, Chapter II.1.1, I have such a desire or pro-attitude toward this object if, when I fail to achieve it, I experience disappointment, frustration or regret.

But that my action is directed toward the achievement of this object does not imply, even minimally, any such pro-attitude it in any nontautological sense. For example, I may be caused to purposefully peel the label off the ginger beer bottle, not by any pro-attitude toward peeling the label off the ginger beer bottle, but rather by anxiety, or habit, or the perception of the dampness of the bottle. If I am prevented from doing so, I may experience neither disappointment, nor frustration, nor regret. Hence that my action is directed toward the achievement of this object does not imply that it is this object that causes me to pursue it. Moreover, even if I did have a pro-attitude toward this object, even this fact would not imply that this pro-attitude is what causes me to pursue it. It is an open question whether it is my pro-attitude toward peeling the label off the ginger beer bottle or my anxiety that causes me to do so. The purpose of an action need not supply its motive.

Of course some purposes of action do supply its motives, as when the intentional object at which my action is directed is one I desire, or aspire, or resolve to achieve. Desires, aspirations, and resolutions are occurrent psychological causes of action that take the agent's purposes as intentional objects and would not be motivationally effective without them. These are the cases in which it makes sense to describe the agent as having a motivationally effective "pro-attitude" toward the purpose of the action. Call these causes of action *forward-looking motives*. Not all forward-looking motives are desires, and not all action is caused by forward-looking motives.

For there are other occurrent psychological causes of action that are unrelated to the purpose of my action, and instead presuppose perceived intentional objects as causes. I considered some of these briefly in Volume I, Chapter VI.5.1 – 2; and at greater length in Chapter V.4.1 above. Here is a further example: perceived traffic jams cause frustration, which motivates honking the horn. Honking the horn is a fully intentional action. I may have a pro-attitude toward honking the horn, but then again I may not. In either case, honking the horn need not be

motivated by its purpose. Instead it may be motivated by an emotion that is caused, in turn, by the perception of an intentional object. Call such psychological causes *backward-looking motives*. My distinction between forward- and backward-looking motives parallels Michael Stocker's distinctions between the "in order to"/"for the sake of" and the "out of"/"from" locutions.[12] My claim is that much action is motivated solely by backward-looking motives.

Backward-looking motives, in turn, may be of three kinds. In the example just described, the immediate psychological cause of action is an emotional reaction to a perceived intentional object. Describe such motivationally effective emotional reactions as *affectively motivating states*. Affectively motivating states constitute one kind of backward-looking motive. But sometimes perceived intentional objects can elicit a goal-directed behavioral response almost automatically, without the intervention of an affectively motivating state, if the disposition to respond to that perceived intentional object in that way is deeply instilled, as when I respond to the perceived ringing of the telephone by picking it up and saying, "Hello?" Call these *perceptually motivating states*. In these cases, the mere perception of an object is motivationally effective in causing an overt behavioral response directed toward a different object. In a comparable manner, the cases discussed in Chapter V.4.4 – 4.5 above are examples of *conceptually motivating states*, in which the mere occurrent thought of an abstract object – a concept, principle, or declarative proposition – is motivationally effective in causing an overt behavioral response directed toward a different, perhaps equally abstract object as its purpose.

Affectively, perceptually, and conceptually motivating states are all species of backward-looking motive. All are genuine motives with identifiable intentional objects, rather than mere whims, impulses or appetites that nonrationally assail us. So they do not fit Nagel's description, discussed in Volume I, Chapter VII.2.3, of unmotivated desire. Yet they are as familiar and plentiful in our experience as any of the various types of desire to which Humeans confine their attention.

7.2. Motives and Respect for Principle

Now according to the prevailing Humean model of motivation, any such backward-looking motive must be followed by a forward-looking motive, namely a desire, if it is to cause action. Thus, for example, the Humean picture implies that my feeling of expansiveness, caused by my having just got a raise, can only indirectly cause me to scatter dollar bills in the street, by first engendering in me a desire to scatter dollar bills in the street. But no such desire, nontautologically construed, is necessary to explain action. I suggested in Volume I, Chapter VI.5 that it is often sufficient that deeply inculcated norms of social behavior simply dispose me to react or behave in certain ways in response to my perception of a situation as being of a certain kind. In the present example, my emotional reaction to getting a raise, i.e. my feeling of expansiveness, is direct in that it is unmediated by any conscious conception of how I ought to feel or behave under these circumstances. And this affective motivational state in turn causes me to perform a purposeful action, namely to scatter dollar bills in the street. But this action is

equally unmediated by any desire or "pro-attitude" toward scattering dollar bills in the street, for I would feel no frustration or regret were I prevented from doing so. (But even if I did, it still would be a moot question whether it was my pro-attitude toward scattering dollar bills in the street, or my expansive feeling, that caused me to scatter them.) My motive for doing so is that I am feeling expansive. And I was caused to feel expansive by having just got a raise.

Thus a backward-looking motive (my feeling of expansiveness) can cause purposeful action (scattering dollar bills in the street) without the intervention of a forward-looking motive. There are other examples of backward-looking affectively motivating states. Free-floating anxiety, consequent on my perceived social incompetence, causes me to roll my napkin into little balls at dinner. Irritation at my government's obtuseness causes me to bang the plates and cutlery while setting the table. Fear, consequent on my awareness that I could be hauled into court by the Internal Revenue Service for income-tax evasion, causes me to pay my taxes.

Similarly, feelings of respect for the moral imperative to aid imperiled friends first (other things equal), consequent on my awareness of Ellsworth as an imperiled friend, causes me to rescue Ellsworth first. An intentional object, i.e. a friend's peril and my prima facie obligation to aid him, causes a backward-looking affectively motivating state, i.e. respect, which in turn causes a purposeful action, i.e. my rescuing Ellsworth first. I feel respect for imperatives thus derived from my moral theory, because – as I argued in Chapter V.5.1 – I feel the force of logic, and also the immediacy of the application of this theory to our situation. My moral theory governs my understanding of the events I perceive – i.e. that Ellsworth is imperiled and that I must rescue him right away; and it motivates my responses to them – i.e. my direct and unambivalent attempt to rescue him. But my moral theory would neither inform my perception of this emergency situation nor precipitate my rescue of Ellsworth, if it furnished no guidance for the treatment of friends, nor for rendering aid to the imperiled. And of course no one would be tempted to take seriously a moral theory as impoverished as that. Only a theory capable of guiding and making sense of moral experience in practice can elicit our respect – or, for that matter, our attention.

However, I could not identify Ellsworth as an imperiled friend relative to my respected moral theory, were it not for my prior, unmediated affection and concern for him. Anti-Rationalists tend to speak as though to have an overriding personal investment in an impartial moral theory is not only to suppose that moral principles apply to all human agents (true), but also to be motivated primarily by concern to conform to the imperatives of this theory to enter into personal relationships in the first place (false). Thus Bernard Williams claims that for the Kantian, "personal relations at least presuppose moral relations. … [T]hey are applications to this case of relations which the lover, qua moral person, more generally enters into."[13] Similarly, Michael Stocker argues about Utilitarianism as follows:

> Suppose you embody this Utilitarian reason as your motive in your actions and thoughts toward someone. Whatever your relation to that person, it is necessarily not love (nor is it friendship, affection, fellow feeling, or community). The person you supposedly love engages your thought and action not for him/herself, but rather as a source of pleasure.[14]

William's and Stocker's criticisms both assume that the moral principles that guide and motivate my action must be in the forefront of my mind as I respond to and interact with others, prompting me to assess and seek out persons and situations that instantiate them. That is, both criticisms assume that *to be guided and motivated by moral principle is thereby to treat others merely as means to obeying it*. These criticisms, so Kantian in spirit but which assume so unquestioningly the Humean models of motivation and rationality, are false.

To be sure, if I am in fact a moral person, then moral principles apply to my personal relations, and my behavior toward others either exemplifies or violates these principles, regardless of any changes in my attitude toward either: All is fair neither in love nor in war. If, further, I identify myself as a moral person, then the principles derived from my moral theory not only apply to my personal relations, but also guide them. But that moral principles apply to and guide my personal relations cannot imply that my personal relations presuppose moral relations. For if we could have no personal relations without presupposing moral relations, there would be no examples for the principles that define moral relations to apply to. If I had not already befriended Ellsworth and recognized his peril, I could not obey the moral imperative to aid imperiled friends first, other things equal, in our situation. And if I bore no such personal relation to anyone, obviously this imperative could have no application at all.

Hence my respect for this imperative need not blind me to Ellsworth's uniqueness, nor pre-empt my friendship for him, any more than my impartial belief that smoking is unhealthy blinds me to the temptation of the cigarette before me, or pre-empts the craving to which I am in danger of succumbing. It is an interesting view of moral obligations that regards them as stifling or distorting our personal relationships; as though the obligation to treat a friend with special care somehow took all the fun out of it.

Moreover, it is precisely my respect for this moral imperative that obviates any doubt or ambivalence that might otherwise cause me to hesitate in deciding whom to rescue first. If I did not respect my special moral obligation, other things equal, to friends, my disposition to rescue Ellsworth first might be overriding, but it would not be unqualified by ambivalence about where my moral obligation lay. Without my recognition of Ellsworth as a friend, my disposition to rescue him first might not be qualified by qualms about my moral duty, but it might not be overriding either. Being motivated to rescue Ellsworth first by this moral principle, then, as much presupposes an unmediated personal relationship to Ellsworth as it does respect for my moral theory.

Some might maintain that it is precisely the potential for ambivalence, or for a conflict between friendship and duty, that shows the fundamental defect of strictly impartial moral principles. That they might prescribe one course of action, and my natural inclinations another, reinforces the alienation that I claim is a straw man. But the problem is then not local to impartial moral prescriptions, but instead common to any morality – indeed, to any prescriptions of any kind that happen to diverge from what I am naturally inclined to do.[15] If we think of a morality as, roughly, a way in which our actions and emotions are or should be regulated by the legitimate

requirements of others, then the objection is, in fact, an objection to heeding those requirements at the expense of one's personal inclinations, and a complaint that one is not invariably encouraged to indulge them. Such a complaint is of course fully in keeping with the egocentric cast of the Humean conception, as well as with Nietzsche's motivational ideal of spontaneity discussed in Chapter V.6.1, above. But I argued in Volume I, Chapter VIII.3.2.4 that such a complaint in the end bespeaks narcissism of pathological dimensions.

A motivationally effective moral imperative, then, ordinarily presupposes rather than precludes unmediated feelings of affection, compassion, or concern. So to be motivated to rescue Ellsworth first by respect for a moral imperative does not imply that my purpose in acting is to obey that imperative to the detriment of my overriding concern for Ellsworth, any more than being motivated by fear of the IRS to pay my taxes implies that my purpose in acting is to obey the IRS to the detriment of my overriding concern to pay my taxes. In both cases, my complex response to a perceived intentional object (the specter of the IRS, a friend's peril) includes a backward-looking affectively motivating state (fear of the IRS, respect for the moral law) that motivates purposeful action (paying my taxes, rescuing Ellsworth first).

7.3. Moral Integrity

So we can think of a *morally integrated agent* as one with a motivationally effective intellect whose moral theory constrains and is fully integrated into the concepts and principles constitutive of her perspective; satisfies the requirements of horizontal and vertical consistency over time and so is rationally intelligible to her; and is such that she recognizes herself in its principles. Furthermore, the character dispositions that these principles describe are sufficiently deeply instilled, preferably in the normal process of socialization, as to reinforce and strengthen the motivational efficacy of her intellect particularly with regard to moral demands. This means that her actions express genuine preferences that are motivated and guided by moral principles in two ways.

First, she naturally develops relationships with others that elicit mutual trust, affection, respect, etc., or their opposites; and interprets these relationships, actions, emotions, and individuals with the help of the strictly impartial vocabulary of concepts and principles her moral theory supplies. Thus she view people's actions, her own included, as right or wrong, well-intended or maleficent, honorable or shameful, and so on; and people themselves, herself included, as accordingly judicious or partial, benevolent or malevolent, virtuous or vicious, innocent or corrupt, generous or spiteful, good or bad, and so on. That is, she recognizes the terms and principles of her moral theory to apply to her experience.

Second, these morally theory-laden judgments reinforce some affectively, perceptually or conceptually motivating states at the expense of others and some behavioral dispositions at the expense of others. Thus, for example, her judgment that she is selfish makes her feel ashamed, and so motivates her to behave unselfishly; her judgment that others are beneficent disposes her to reciprocate; her judgment that another is suffering makes her feel compassion, and so moves

her to render aid; her judgment that injustice is being done moves her to right it. That is, her morally theory-laden experiences reinforce or undermine her moral training. In Chapter VIII.6 below I elaborate this conception of moral integrity at greater length, and in Chapter IX below I say more about how certain morally theory-laden experiences might undermine an agent's moral training.

On the above account, it would be misleading to deny that an agent has a conscious commitment to his moral theory; for its concepts and principles saturate his interpretation of morally appropriate behavior, of his own emotions and actions, and of himself and other people. He thinks of them as, for example, friends, responsible agents, rational beings, loved ones, etc. and responds to them accordingly. But it would be similarly misleading to object, as Blum, Williams, and other Anti-Rationalists such as Wolf and Stocker do, that his moral theory alienates him from the objects of his moral concern. For it is only with the aid of his moral theory that he is able to recognize situations as being those in which compassion, for example, is appropriate. Without his moral theory, he would lack the concept of a person as good, valuable, a friend, or deserving of aid or respect. Without these concepts, it is unclear what would cause him to feel compassion for her.

We have seen that compassion presupposes empathy and sympathy with another's inner states, and that both presuppose our ability to modally imagine those states to ourselves – and so to interpret and identify conceptually the other's inner states as states of pain or suffering; indeed, pain or suffering to which empathic and sympathetic responses are appropriate. Without this prior, theory-laden conception, we would have no way to know what empathic and sympathetic inner states of our own might correspond to the other's, and thus no cognitive basis for the felt psychological connection with the other that a visceral understanding of her interiority requires. Then it is equally unclear what would motivate our moral behavior toward him. In order that we experience the direct moral emotions – such as compassion – on which Humeans and Anti-Rationalists insist, a strictly impartial moral theory that saturates our modal imagination of others' interiority must be presupposed.

8. Explaining the Whistle-Blower

Now to recur to the problem raised in Volume I, Chapter VI.5.2, of how to explain the whistle-blower's actions, if not in the familiar Humean terms. Recall that we acknowledged that some whistle-blowers' actions might be explicable in these terms;[16] but that most probably were not because the projected satisfactions of seeing justice done were so rare, uncertain or bittersweet at best. The majority of whistle-blowers cited different reasons for making public their employer's harmful or unethical workplace and/or business practices: disgust or outrage with others' arrogance and dishonesty;[17] a belief in open information, truth, justice, or reason;[18] loyalty to the public;[19] conscience or personal ethical or religious principle;[20] a sense of personal responsibility or obligation to others[21] – even though they had everything to lose and nothing to gain. Many whistleblowers felt that they had no choice, that they were forced or compelled to

expose the corruption of their organizations.[22] We also looked at the explanation offered by Socrates, the most famous whistle-blower of them all, for exposing the ignorance and pretentiousness of his fellow citizens and refusing to retract his criticisms even though his stubbornness condemned him to death:

> 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. … it 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.[23]

The whistle-blower, then, is the direct antithesis of the free rider discussed in Chapter IV.8 above. Whereas the free rider violates moral rules for the sake of personal advantage, the whistle-blower violates immoral rules for the sake of others' advantage. Whereas universalization of the free rider's behavior leads to social chaos, instantiation (or, for that matter, universalization) of the whistle-blower's behavior leads to social benefit. And whereas the free rider's success depends on private concealment, the whistle-blower's success depends on public disclosure. In essence, the whistle-blower redresses the injustices free riders collectively perpetrate. In this respect, the whistle-blower is a significant though widely neglected paradigm of right conduct for moral philosophy.

The foregoing discussion supplies the necessary apparatus for a character profile of the whistle-blower that proposes to explain how she might be motivated to redress injustice even when desire plays no role. Three predominant traits that characterize the whistle-blower are genuine preference, interiority, and motivationally effective intellect. First, the whistle-blower is consistent, or asymptotically approximates consistency, in her choices over time, in the sense discussed in Chapters III and IV. This is a case that shows how my and McClennen's psychologies of choice are mutually interdependent: the whistle-blower remembers and is guided by the priorities she established in earlier choices as she is choosing on the present occasion; chooses in the knowledge that she is similarly accountable to the future self that reflects back on this present choice; and so commits herself resolutely to that course of action on which she knows she can follow through. That is, the whistle-blower knows going in that she will not chicken out. This is what it means when a whistle-blower says that she could not have lived with herself, had she not taken the action she did. In this comment she expresses her awareness that she is a consistent self that persists through time; that she must at each moment carry forward the implications and consequences of choices she has made in the past; and that she must now

choose in such a way as to be able to justify that choice to the self she will later become as the result of them. My psychology of choice describes this awareness; McClennen's psychology of choice describes the overriding value the agent consciously ascribes to it. This is the awareness Socrates expresses when he says,

> Wherever a man's station is, … there it is his duty to remain and face the danger … it would be very strange conduct on my part if I were to desert my station now … when the god has commanded me … to spend my life in searching for wisdom, and in examining myself and others.

To say that an agent places an overriding value on consistency of choice through time, i.e. on acting on her genuine preferences, is to say that she is committed to resolute choice as a principle that regulates the choices she makes at each moment in time.

Second, therefore, the whistle-blower has a highly developed interiority. Feelings of fear, greed, or self-seeking are controlled, suppressed, and ultimately outweighed by the vividness and intensity of her perception of the injustice; by that of her symmetrical modal imagination of the harm to others this injustice does; by the force of the moral emotions she experiences in response; and by the force of the impartial, normative moral or religious concepts and principles that saturate her perceptions, precipitate and systematize her responses, and chart the particular course her intertemporally consistent choices take. These are the causal factors that decisively effect the whistle-blower's action. Her action is a response to these gripping interior cognitive states and events – not to any desire to maximize her utility or impulse to cave in to external pressure. This is the response that Socrates defends as an overriding value when he says,

> 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.

In this passage Socrates not only expresses his overriding commitment to the demands of his own interiority in acting as he has; but argues that this priority – of the interior demands of conscience and principle over external threats or inducements – should be overriding for "a man [or woman] of any worth at all."

Third, therefore, the whistle-blower has a motivationally effective intellect that marshals and organizes her perceptions, modal images, and emotions under the normative moral or religious concepts and principles that define her perspective and saturate her character dispositions. These concepts and principles turn the raw, sensory data of her perceptions and reactions into meaningful experiences. These, in turn, reinforce and underscore the causal efficacy of the principled dispositions that motivate and guide her actions. For Socrates, the overriding principles were justice, virtue, self-analysis and the search for wisdom. Other principles cited by whistleblowers include honesty, humility, truth, justice, reason, loyalty to the public, and altruism. The final measure of the motivational efficacy of these principles is in the actualized dispositions to action that instantiate them. In the end no other measure, including verbal allegiance, is important.

Now one of the traits of the whistle-blower that most stymies attempts to explain such behavior in Humean terms is his capacity to resist the pressure of his peers to remain silent and conform – to "go along in order to get along." Of course the whistleblower's resistance to the pressure of morally indefensible conformity can be reduced to no more than a particularly interesting instance of sacrificing self-interest in order to act on moral principle. But it is more than that. The whistle-blower's resistance involves not only personal sacrifice and discomfort, but also a willingness to sacrifice the intersubjective interpretation of his moral obligations that receives affirmation and validation from a consensus shared with others to an interpretation that may be affirmed, validated and shared with no one other than himself. That is, it involves a willingness to come into direct conflict with the actual moral community whose shared practices have in the past given life and physical substance to the moral principles on which he now acts; this is the moral motive that, as we saw in Volume I, Chapter IX.1.4, Anderson's Noncognitivist model of moral justification could not accommodate. The whistle-blower's resistance to social pressure thus involves sacrificing that part of himself that most directly connects him to those others whose actions helped to instill and reinforce the principled dispositions he in this instance upholds. In order to uphold the moral principles his actions embody, he must reject the counsel – and the demands, and the pressure, and the bribes, and the threats – of those in whom he has previously found validation for them; and face the punishing consequences of doing so, in addition to the punishing consequences of blowing the whistle itself. In order to detach himself from the consensus opinion of his surrounding moral community, and so leave behind that part of himself that identified himself as part of it, he must distinguish sharply between the moral conviction that arises solely from his own, rational evaluation of the situation; and that which arises as the result of interpersonal dialogue with and mutual affirmation by others. He must distinguish between these two, and reject his morally supportive relationship to others – to friends, to family, to community, to peers. This is no easy thing to do.

What enables the whistle-blower to do this – and so what distinguishes him from the vast majority of his peers – is precisely the possession in preponderance of the above three traits: genuine preference, interiority, and motivationally effective intellect. In Chapter IV.8 we saw that an agent's violation of her commitment to resolute choice called forth sanctioning moral emotions – guilt, shame and resentment – directed by herself toward her own moral dereliction; and that violations of intertemporal consistency were sufficient to inflict on herself such sanctions even in a putatively non-moral case like breaking a diet. With the aid of the concepts of interiority and strict impartiality, we can now see that in fact there is no purely nonmoral case and no purely nonmoral action: An agent is always answerable to himself, his memories and his emotions for the intertemporal consistency of every action he performs – for its present consequences in his life; the interior recollections, responses and habits it instills in him; and the expectations and attitudes about the future it implants. The symmetry requirement of strict impartiality that holds between the self and the other holds between the self at one time and the

self at another as well. This is the essence of Nagel's and McClennen's – and my – rejection of pure time preference as a principle of rational choice.

So in order to act as she does, the whistle-blower must draw not on her previous bonds with others and the support she derived from her interactive relation to them; but rather on her own, constantly evolving bonds with her own, earlier incarnations and the support she derives from her interactive relation to them. These bonds of mutually reinforcing action over time, according to principles that regulate all such act-tokens as they occur, forge a strong and internally consistent psychological foundation of entrenched dispositions of character that fortify her interior moral conviction against the disapproval, rejection and retaliation she experiences at the hands of that external moral community she must now disown.

We all like to think we would do the right thing if placed in a situation in which we were forced to choose between morality and self-interest – that is, in which we could do the right thing only at the cost of considerable inconvenience, sacrifice or danger to ourselves. But in fact these situations never present us with choices in the ordinary sense of that word. Rather, they offer us the opportunity to find out which of the guiding concepts and principles we espouse are in fact motivationally effective for us, what our true priorities are, what we are made of, and who we truly are. We are all offered ample opportunity for such self-knowledge on a daily basis; in Chapters VII and VII below, I anatomize the resourceful strategies by which we often evade it. For in fact any agent presented with those same opportunities reacts in a way analogous to that of the whistle-blower, regardless of how he chooses: When the whistle-blower says of herself that she "had no choice," or was "forced or compelled" to act, what she is really saying is that given all of the causally effective elements that combine to constitute her character, acting in character was the only choice she had.

Endnotes to Chapter VI

[1]Lawrence Blum, *Friendship, Altruism and Morality* (Boston: Routledge and Kegan Paul, 1980), 3. Henceforth all page references to this work will be parenthecized in the text.

[2]I make this concession to non-Kantians only because considerations of space preclude more extended argument to the effect that without modal imagination and *bona fide* concept-formation we would have no first-personal access to our motives, thoughts or emotional states at all. Nothing of consequence for my argument turns on this concession.

[3]How should we analyze our feelings towards the masochist? This depends on the correct description of masochism. If masochism involves feeling pleasure in response to an experience that would cause us pain, then it may difficult to empathize with the masochist's inner state, since difficult to viscerally understand it; more difficult still to sympathize with his inner state, since difficult to for us to feel concordantly; and impossible to feeling any *immediate* inclination to render aid since, according to this description, he does not suffer. So whatever we may feel about this brand of masochist, it will not be, on this account, compassion. Of course we may feel distressed or shocked that he takes pleasure in what causes us pain, and feel inclined to try to reform him. But this would be paternalism at best, meddling at worst.

Suppose, however, that the correct description of the masochist is that he takes pleasure in his own pain; i.e., that he experiences two opposing states, consecutively or simultaneously, where we would feel only one, namely pain. Then we might both empathize and sympathize with his pain, and also feel an inclination to render aid – an inclination that is, however, dampened by our recognition that, astonishingly, he would prefer none. In this case I think we should simply say that we feel compassion compounded by incomprehension, frustration, revulsion, and so forth.

[4]Could one be impartial in one's judgment if both one's own and the other's interests were equally surface, rather than depth objects of imagination? Since symmetry would remain inviolate, why not? Since, in this case, one's capacity to understand any of the interests in question would be vitiated, *a fortiori* one's capacity to judge them impartially would be as well.

[5]For example, consider the California association of African-American social workers that has successfully lobbied for legislation prohibiting the adoption of African-American children by Euroethnic families, even when those families have served the child in the capacity of foster parent for a sufficiently extended period of time that strong emotional and psychological bonds have formed between foster parents and child. The association's reasoning is that African-Americans in general are best served by being raised in cohesive African-American families – a concern with which all adult African-Americans can identify. What the association seems to lack is the empathic understanding of what it means to a child to have psychological bonds of trust and affection with an adult caretaker destroyed, and destroyed repeatedly as the child is moved from one foster home to another; and what toll this will take on the child's capacity to form bonds of trust and affection with anyone as an adult. It would seem that the association's failure of

imaginative involvement with the child's inner states as depth objects, and correspondingly deep imaginative involvement with the long-term interests of adult African-Americans as a group, incapacitates its members from impartially carrying out their mandate to protect and promote the child's best interests.

[6]Obviously this assumption becomes more problematic as the number of empathees increases. Possibly some adaptation of the method of pairwise comparisons might be useful here.

[7]Rawls reconstructs this view from Hume and attributes it to classical utilitarianism in *A Theory of Justice* (Cambridge, Mass.: Harvard University Press, 1971). See pp. 33 and 184-7, and also 27-8. Also see Lawrence Kohlberg, "The Claim to Adequacy of a Highest Stage of Moral Judgment," *The Journal of Philosophy LXX*, 18 (October 25, 1973), 630-646.

[8]Thus Rawls' own view is that impartial judgments are those that result from observing the conditions characterizing the original position, especially the veil of ignorance (of one's own interests and position in society). See Volume I, Chapter X for extended discussion of Rawls' view.

[9]The connection between impartiality and compassion is particularly evident in Hume's *Treatise of Human Nature*, Ed. L. A. Selby-Bigge (Oxford: Clarendon Press, 1974). See the first parts of Book II, Part I, Section XI, and Book III, Part I, Section I; and also Section VI.

[10] Although to point this out is not necessarily to justify the theory, or to account for its origins.

[11] This distinction is first made explicitly by Kant, in *Kritik der Praktischen Vernunft*; see 2C, Anmerkung I to Lehrsatz IV. H. A. Pritchard relies on this distinction, although he uses it to different ends, in "Does Moral Philosophy Rest on a Mistake?" *Mind XXI*, 81 (January 1912), 21 – 37.

[12] See Stocker's "Values and Purposes: The Limits of Teleology and the Ends of Friendship," *The Journal of Philosophy LXXVIII*, 12 (December 1981), 747 – 765.

[13] Bernard Williams, "Person, Character and Morality," in *Moral Luck* (New York: Cambridge, 1981).

[14] Michael Stocker, "The Schizophrenia of Modern Ethical Theories," *The Journal of Philosophy LXXIII*, 14 (August 12, 1976), 458.

[15] Marcia Baron also makes this point in "The Alleged Repugnance of Acting from Duty," *The Journal of Philosophy LXXXI*, 4 (April 1984), 213.

[16]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 ?.

[17]Glazer and Glazer, *ibid.*, page 19, 100., 122, 138, 223, 246. Also see Mary Schiavo, "Flying into Trouble," *Time* (March 31, 1997), pages 52-62.

[18]*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).

[19]*Ibid*., pages 17, 40, 45, 129.

[20]*Ibid*. pages 43, 70, 88, 96, 101, 103, 104-5, 117, 119, 122, 141, 248-9. Also see Clyde H. Farnsworth, *op. cit.* Note 16; and "In Defense of the Government's Whistle Blowers," *The New York Times* (Tuesday, July 26, 1988), page B6.

[21]*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.

[22]*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.

[23] 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.

PART TWO: REALITIES

> Now reason enjoins its prescriptions relentlessly, without holding out any prospect to inclination; therefore, so to speak, with disregard and neglect of these impetuous and therewith so seemingly humble claims (which refuse to be subdued by any command). From this there arises a natural dialectic, that is, a propensity to pseudorationalize [*vernünfteln*] these strict laws of duty – to call into doubt their validity or at least their purity and rigor, and where possible to make them more accommodating to our wishes and inclinations; that is, basically to corrupt them and destroy their entire dignity, which in the end even ordinary practical reason itself cannot approve. [G, Ak. 405]

With the ideals of transpersonal rationality, consistency, and moral motivation now in place, I turn next to extended discussion of the non-ideal realities in which these ideals serve – psychologically, morally and socially – as distant reminders of the standards of performance to which we naturally aspire. This second part of the discussion recapitulates and develops analysis of the conflict between transpersonal and egocentric rationality detailed in Chapter I – with closer attention now to the practice of philosophy itself on which that Chapter focused – by extending further the exploration of motivationally ineffective intellect begun in Chapter V.4.2.1. That analysis of moral motivation was placed in Part I because it disregarded familiar, real-world impediments to enacting the dictates of reason, including moral dictates. Part II takes those impediments for granted, and focuses on the downward cognitive and psychological accommodations we make to the reality of our rational and moral insufficiencies. Just as the concepts of horizontal and vertical consistency are key to understanding how reason guides action under ideal circumstances, the concepts of literal self-preservation introduced in Chapter V.2, and of pseudorationality introduced in Chapter VII below, are key to understanding how reason guides action under actual circumstances. So they figure prominently in subsequent chapters.

Chapter VIII.4 surveys selectively some of the accounts of pseudorationality furnished by the history of philosophy, of course with particular attention to Kant. All of these historical accounts identify systematic deviations from an ideal of rational integrity that is implicit in classical logic and presupposed by Aristotle, Kant and Nietzsche. But all of the accounts themselves are scattered and cursory rather than systematic. Consequently, I intend my account of pseudorationality first to consolidate and build on some of these historical accounts; and second, to offer very tentative preliminary guidelines for more systematic, in-depth investigation of rule-governed deviations from theoretical rationality, analogously to the way in which Tversky and Kahnemann's work provided the foundations for investigating rule-governed deviations from instrumental rationality.

In the case of theoretical rationality, the primary act of cognitive distortion is the distortion of theory itself – whether of ourselves or of reality – in order to defend it against

the many anomalies that threaten its explanatory sufficiency. The passage above from the *Groundwork* describes one kind of anomaly: a first-personal, internal object of desire or impulse that is inconsistent with the moral theory in which we purport to recognize ourselves, and to which we therefore accommodate our theory through rationalization, i.e. by warping its scope of application so as to include it. A second is the type of anomaly described in the first *Critique* at A 112: a third-personal, external object of perception that is inconsistent with the conceptual scheme that structures the unity of the self, and that we therefore exclude through denial or dissociation. In this case, by thus warping the scope of application of our theory, we thereby warp the structure and crush the healthy functioning of the self, engendering internal blindness, bias, and fear.

These two types of anomaly can be distinguished conceptually, but they work in tandem when one's personal investment in a theory is very deep. This is the case that forms the primary target of the following discussion, because our personal investments in our favored moral theories are, in fact, often that deep. I try to show how the dynamics of this case can explain a wide variety of moral dereliction, and also why we respond so sluggishly to the imperative to ameliorate it. Following this systematic and rather extensive cataloguing of our moral failings, however, is a reconsideration of Kant's account of theoretical reason in the Dialectic of the first *Critique* that leads finally to the conclusion that the very same intellective capacities that inevitably propel us into the many vices of pseudorationality also contain very powerful resources for reforming them.

Chapter VII. Pseudorationality[1]

In Chapter V.2 I condensed the arguments of Chapters II and III in the concept of *literal self-preservation*, i.e. preservation of the horizontal and vertical consistency over time of the experiences, including the genuine preferences, constitutive of an agent's perspective; and so preservation of the rational intelligibility of those experiences. I argued that literal self-preservation – essentially, preservation of theoretically rational coherence – was a necessary condition of rational agency, and motivationally overriding in the structure of the self. In this chapter I show how the structural and motivational dominance of literal self-preservation explains why, when confronted by conceptual anomaly, we are more inclined either to suppress it from consciousness altogether, or to distort or truncate the concepts constitutive of our perspective in order to accommodate it. By thus systematically warping our theory-laden understanding with the collusion of our rational capacities themselves, we achieve the illusion of horizontal and vertical consistency over time, and so of rational intelligibility, against the reality of interior disintegrity. I describe this systematic process as one of *pseudorationality*.

Pseudorationality simultaneously violates the consistency requirements of rational intelligibility, and, in so doing, indirectly confirms their centrality in the structure of the self. It is therefore central in understanding the functioning of transpersonal rationality in the non-ideal case. This is the case in which the vastness and complexity of reality, plus the weight and incorrigibility of our deeply ingrained desires and impulses drag down and thwart our even more deeply ingrained but ultimately futile disposition to soar above and master these impediments to intellective freedom. Anchored and trapped in the gravitational pull of a conceptually overpowering, unmanageable and insubordinate universe, both exterior and interior, we pay homage to our intellective aspirations to transpersonal rationality by furtively adapting it to the ungovernable and overwhelming actual circumstances we originally invoked it in order to master. This adaptation is both necessary and tragic, betraying, as it does, the transpersonal capacities with which we are fitted, by attempting to put them to work in situations – the human situation – that warp and stunt them in the service of desire. Pseudorationality is the expression of our attempt to save face before our humiliating intellective defeat at the hands of the real.

On my account, pseudorationality comprises three mechanisms for coping with conceptual and theoretical anomaly that are implicit in classical logic: denial, dissociation, and rationalization respectively. Briefly, *denial* is a form of biased nonrecognition that degenerately satisfies the comprehensive requirement of rational intelligibility. *Dissociation* is a form of biased negation that degenerately satisfies the requirement of horizontal consistency. *Rationalization* is a form of biased predication that degenerately satisfies the requirement of vertical consistency. There are probably other pseudorational mechanisms besides denial, dissociation and rationalization. But I believe these three to be primary. In all three cases the bias is toward the appearance of rational coherence against the reality of theoretical insufficiency and so of

thwarted literal self-preservation. And in all three cases the satisfaction of rationality requirements is degenerate because it relies on arbitrary and *ad hoc* improvisations that sacrifice the spirit to the letter of those requirements. All three mechanisms are interdependent and mutually supporting points on a continuum of theoretical *ir*rationality, with denial at the pathological extreme and rationalization a merely elaborate demonstration of intellectual agility that nevertheless fails to wrestle with the facts. It is because these mechanisms preserve the appearance of unified selfhood and agency against the fact of disintegrity and the threat of disintegration that I refer to them as *pseudorational* mechanisms.

Section 1 describes these three mechanisms in outline, and offers a brief example to illustrate how they function. Section 2 distinguishes between two kinds of cognitive anomaly identified by Kant that activate these mechanisms – conceptual anomaly and theoretical anomaly – and formulates the cognitive challenge that these present to our attempt to make both the world and ourselves rationally intelligible. This section also makes a further distinction that bisects this one, between first-person and third-person anomaly. Section 3 introduces a playful and frivolous test case of third-person conceptual anomaly, by way of dissecting the disruptive operations of denial on our attempts to integrate conceptual anomaly into a rationally intelligible perspective; and describes in detail some of the cognitive and psychological problems conceptual anomaly poses even under these relatively benign circumstances.

In Section 4 I turn attention on theoretical anomaly, and offer a brief taxonomy of psychological characters – the naïf, the ideologue, the true skeptic, and the dogmatist – for some of whom pseudorational denial may usefully function. I argue that pseudorationality in response to theoretical anomaly bespeaks a personal investment in one's favored theory that places it at the service of desire-satisfaction. Section 5 focuses in more narrowly on the mechanism of denial in response to theoretical anomaly; and offers a criterion for distinguishing rational from pseudorational denial. I focus henceforth on the character of the dogmatist for its probable relevance to all or most of those potential readers of this volume, including me. Section 6 is devoted to discussion of the pseudorational mechanism of dissociation in response to theoretical anomaly, with application to several examples, including that of art-critical responses to contemporary art; and compares my account with Philip Bromberg's. Section 7 addresses the pseudorational mechanism of rationalization, and applies it to the anachronistic racist theories of Jensen and Murray. Section 8 invokes all three mechanisms in the analysis of a real-life historical example. I argue that in the case of first-person theoretical anomaly, they function in tandem as a form of self-deception. In subsequent chapters I explore the implications of this form of self-deception for the possibility of moral integrity.

1. Three Pseudorational Mechanisms

First, a general sketch with details to come later. We saw in Chapter II that vertical consistency requires that the various components of our experience be integrated and unified under the rubric of more general, comprehensive, and motivationally effective concepts and

principles; and that all of these also satisfy the requirement of horizontal consistency relative to one another. For example, take the relatively general and motivationally effective cognitive principle that we are to understand an external event in the world by seeking out its causal relations. This principle is horizontally consistent with that of understanding internal mental events, such as beliefs and feelings, by seeking out their causal origins in our upbringing, social environment, and previous experiences. But it is also similar in its reliance on causal explanation. The more general principle with which both are vertically consistent is that we understand all the phenomena of experience by seeking out their causal connections.

However, there are anomalies to which this more general principle seems not to apply. Then we resort to pseudorational mechanisms in order to explain them. A familiar example is the micro-phenomena studied by quantum physics, which seem peculiarly resistant to causal explanation. Our instinctive response is to begin by *denying* the phenomenon, and to cast about for flaws in the experimental design or apparatus to account for the apparent illusion. The intractability of the phenomenon to our attempts to wish it away are then met by *rationalization*: We argue that there must be a causal explanation of this phenomenon, but that we are insufficiently equipped to discover its causes. When the evidence indicates the untenability of this position, we shrug our shoulders and proceed to *dissociate* the phenomena of quantum physics from the comprehensible world of causal relations we aspire to grasp. We then suffer the perplexity of trying, and failing to see how the principles of quantum physics might be made to fit with everything else we think we know.[2]

We can think of these three self-defensive mechanisms, then, as ways in which our highest-order disposition to literal self-preservation rallies, valiantly but ineffectively, to the challenge posed by conceptually unmanageable anomalies. In *denial*, we suppress awareness of an anomalous particular, property or state of affairs, by failing to recognize it as instantiating concepts supplied by our unified conceptual scheme. In denial, the particulars lost to conscious recognition are those which have a rational claim on it; thus I describe denial as *biased* nonrecognition. I say more in Section 5 below about how to identify such bias. Denial degenerately serves to maintain the overall rational intelligibility of an agent's perspective, by eliminating anomalous particulars that violate its horizontal and vertical consistency; and thereby serves literally to preserve the internal unity of the self.

In *dissociation*, by contrast, the anomaly is not banished from awareness entirely, but rather identified solely in terms of the negation of some subset of the concepts that constitute the agent's perspective; thus is the horizontal consistency of experience degenerately preserved. I describe dissociation as *biased* negation because in dissociation, the use of conceptual negation is driven not by disinterested and impersonal conceptual analysis, but rather by that same highest-order disposition to preserve the horizontal consistency of the agent's perspective – and so, at all costs, the rational coherence of the self – against the delinquent reality that threatens them.

Finally, *rationalization* consists in biased predication; in applying a higher-order concept too broadly or too narrowly to something, ignoring or minimizing properties of the thing that do

not instantiate this concept, and magnifying properties of it that do. The requirement of vertical consistency is thereby degenerately satisfied, by stacking and subsuming higher-order, increasingly malformed concepts pulled and squeezed and stretched and trimmed to perform the explanatory task of the moment. Think of these three mechanisms, then, as parts of a mental demolition process, whereby the rational scaffolding of reality is gradually dismantled, so that a funhouse monument to wishful thinking can be erected in its place. Thus does our highest-order disposition to literal self-preservation buckle under stress.

2. Conceptual vs. Theoretical Anomaly

Degrees of conceptual unmanageability are relative to the limitations of an agent's perspective, and different agents' perspectives are limited or inclusive – i.e. provincial or cosmopolitan – to different degrees. The extent to which an agent's perspective is limited or inclusive, and so the degree to which a phenomenon is anomalous or recognizable to her respectively, depends on the agent's interests, background, experience, environment, available information, and opportunities for obtaining more. The natural sciences identify certain phenomena on the anomalous nature of which almost anyone trained in the Western tradition can agree. But that even these are not necessarily anomalous to every human perspective, and that other phenomena similarly may function as anomalous relative to some agents' perspectives but not others, deserves emphasis.

We exacerbate the challenge posed by conceptual anomaly and multiply its instances through incuriosity, by choosing to know and experience less. The less we explore and inquire, the more our minds shrink and grow shallow, rigid, fragile and scared, because the more easily disturbed and disoriented by virtually any trivial curiosity: The sight of a student sporting dreadlocks, or an interracial couple, or a Scotsman wearing a kilt all become threats to literal self-preservation that therefore must be denied, dissociated or rationalized. On the other hand, we master the challenge posed by conceptual anomaly and reduce its instances through curiosity and inquisitiveness, by actively seeking to learn and experience more. The more variegated the types of experiences and ideas the system of concepts and principles constitutive of our perspectives can accommodate, the bigger our minds become and the deeper their interiority; and the fewer the number and magnitude of conceptual anomalies we are likely to encounter. Our highest-order disposition to literal self-preservation is strengthened, not stressed, by a receptive curiosity about the unfamiliar. I recur to this thesis in Chapter XI; and in Sections 4.1 – 4.4, below, describe in greater detail a variety of attitudes that increase or decrease our propensity to pseudorationality.

Nevertheless, there are two distinct kinds of pseudorationality, corresponding to two different pronouncements Kant makes regarding the necessary conditions for the unity and integrity of the self. The most basic is expressed by Kant's assertion in the first *Critique* that

> without [the synthetic unity of appearances according to concepts], which has its a priori rule, and subjects the appearances to itself, no thoroughgoing and universal, therefore

> necessary unity of consciousness in the manifold of perceptions is to be found. These [perceptions] then would not belong to any experience, therefore would be without an object, and nothing but a blind play of representations, that is, less even than a dream. [1C, A 112]

Kant here describes the coherent conceptual organization and systematization of representations that a unified self presupposes, and adds that any representation that fails these conditions cannot enter into conscious experience at all. These are the conditions I attempted to capture in Part I under the general rubric of rational intelligibility; and I suggested in Chapter II.4.4 that this passage introduces into Western philosophy the possibility of an unconscious, in which representations may be deposited – and may have causal efficacy – even though they form no part of an agent's conscious experience and therefore remain unrecognized within the parameters of her perspective. I shall continue to describe as a *conceptual anomaly* an event, object or state of affairs that fails to satisfy the consistency conditions enumerated in Part I.

Such an event, object or state of affairs can be third-personal, in case it issues from an external source, as in the quantum physics case; or first-personal, in case its source is internal to the agent. A sudden and overpowering mood change that is discontinuous with the agent's personality; or the spontaneous appearance of behavior, interests or tastes that are in discord with the agent's settled character; or the unexpected manifestation of a capacity or limitation might exemplify first-personal conceptual anomalies. In any such case, the event, object or state of affairs would count as anomalous even relative to the most comprehensive and cosmopolitan perspective; but less unfamiliar ones can also count as genuinely anomalous relative to more constricted and provincial perspectives. In all such cases, the pseudorational mechanisms of denial, dissociation and rationalization function to preserve at least the semblance of rational integrity in the self against the threat of disintegration such a phenomenon presents.

I called attention to the second kind of pseudorationality in the passage from the *Groundwork* that opened Part II of this volume. This is the case in which a first-personal anomaly, a delinquent inclination within the agent that fails to cohere with his interior self-conception, gives rise to the pseudorational process of rationalization specifically. As explained in Chapter I, a *self-conception* is a theory that includes the properties an agent thinks accurately describe him psychologically, socially and morally, and the more complex principles he thinks govern his behavior and relations with others at a given moment: his attitudes, beliefs, emotions, actions, and consequent image in his own and others' eyes. An agent's desires are determined by his self-conception because they are determined by what he conceives himself as lacking. Here the agent butchers the requirements of particular rational and moral principles in such a way as to provide a justification for indulging those desires at their expense. This is Kant's definition of evil, in which we subordinate the requirements of principle to the demands of desire-satisfaction rather than the other way around [R, Ak. 36], then rationalize this by minimizing the authority of principle and magnifying the value of desire [R, Ak. 42]. As Baron points out, this brand of rationalization naturally leads to externalism about moral motivation:

> [I]nsofar as I adopt a policy of seeking assistance from the inclinations, I in effect allow that moral considerations are not fully compelling, motivationally; but if they are not fully compelling motivationally, presumably they are not fully compelling, period. I will end up asking myself, 'It's obligatory, and that carries some weight; but what other reasons are there for doing it?' (155)

Since such principles enter into the constitution of one's interior self-conception, a morally anomalous impulse or desire that violates the agent's moral theory is analogously a threat to its rational integrity. As Kant so clearly saw, desire exerts a gravitational effect on the intellect, pulling it off center and biasing it towards the satisfaction of that desire. Morally anomalous desire has the same effect, but does even greater damage, by not only biasing the intellect toward its satisfaction, but in addition inciting the intellect to fabulate a self-deceptive rationale for doing so. In what follows I refer to an event, object or state of affairs that is anomalous relative to one's theory of oneself or the world as a *theoretical anomaly*. Theoretical anomaly is a species of conceptual anomaly. I define it further in Section 4, below.

The relation between such a desire and the agent's morally inflected self-conception is analogous to the relation between a first-person conceptual anomaly, such as a sudden surge of aggression in an otherwise placid character, and the coherent conscious experience that it disrupts. Just as denial functions to relegate such an impulse to the unconscious in the latter case, similarly it functions to relegate desire to the subconscious in the former. Whether conceptual or theoretical, the anomaly thus denied may continue to affect the agent's intellect, behavior and perspective – without, however, receiving the conscious recognition that would integrate it as rationally intelligible within that perspective or theory respectively.

However, we have seen in Chapter V.2.2 above that a rationally coherent self is a necessary condition for even having a desire, so the analogy can never be an equivalence. A conceptual anomaly is a threat to the rational integrity of the self and to the rational intelligibility of an agent's perspective; whereas a merely theoretical anomaly puts pressure primarily on the agent's morally inflected self-conception or theory of the world. How closely or loosely entwined this is with the agent's perspective, hence how closely or loosely theoretical anomaly is with conceptual anomaly more generally is discussed in Section 4 below. So although morally delinquent desire may disrupt that self-conception and call forth the mechanisms of pseudorationality to perform repairs, it does not *necessarily* undermine the coherence of the self whose desire it is. An agent's morally inflected self-conception can be fully destroyed by her own anomalous and delinquent impulses without necessarily destroying the rational integrity of the self whose self-conception it is. This is one reason why it is not possible to derive any particular normative moral theory directly from the rational criteria of a genuine preference. An agent whose morally inflected self-conception has been destroyed or debased by her own delinquent desires is very dangerous; I discuss this case further in Chapter IX. But this does not entail that she is irrational.

As indicated above, both conceptual and theoretical anomalies can be of two kinds. *Third-person conceptual anomalies* are events, objects or states of affairs in the external environment that violate the horizontal and vertical consistency over time of the concepts constitutive of our perspective, and so the conceptual presuppositions by which we make that environment rationally intelligible. *Third-person theoretical anomalies* are events, objects or states of affairs that violate the horizontal and vertical consistency over time of our theories about the world or external environment.

First-person conceptual anomalies are those events, objects or states of affairs in the interior environment – our own behavior or emotional reactions, or thoughts or psychological attitudes or states of mind – that similarly violate the horizontal and vertical consistency over time of the concepts constitutive of our perspectives on ourselves as interiorized agents, and so the conceptual presuppositions by which we make ourselves rationally intelligible to ourselves. *First-person theoretical anomalies* are those events, objects or states of affairs that violate our theories of ourselves. Since I ultimately wish to address issues pertaining to normative moral theory, the following discussion will be concerned primarily with the type of pseudorationality called forth by first-personal theoretical anomaly; I embark on this exploration in Chapter VIII. But it will be convenient to dissect the operations of pseudorationality by beginning with third-person conceptual anomaly, because it is usually easier to see things at a distance from oneself, and sometimes easier to see them than to theorize about them.

3. Test Case #1: Encounter on West Broadway

Suppose, then, that you are in New York, making your way down West Broadway, where anything may happen, and you suddenly encounter – what? It is large, mottled gray, prickly, shapeless, undulating, and it moos at you. You have at the disposal of your current perspective certain concepts of higher-order properties that might enable you to recognize this entity – street sculpture, advertising gimmick, genetic mutation, three-martini lunch hallucination, tropical plant, etc.; but it is not immediately evident which one would suffice in these circumstances, or whether any of them would. It is tempting to think that this is just the sort of case that belies the necessity of the requirement of vertical consistency to rational intelligibility. For in this case, it may seem, you must know at least that you have encountered a gray blob, even though you don't know what higher-order kind of gray blob it is.

But reconsider. If it is unclear which of those higher-order properties now at your disposal would enable you to recognize this entity, any of them might. If it is unclear whether any of them would, none of them might. If none of them did, concepts that would enable you to recognize this entity would not form part of your current perspective. I am using the term "recognition" here in the technical Kantian sense, spelled out in Chapter II.3, of *recognition in a concept*, to denote subsumption of some lower-order particular under a higher-order concept that classifies it along with others of its kind, and thereby renders it accessible to unified consciousness. Because recognition in this Kantian sense applies to subsentential constituents of

sentences, rather than to the sentences in which they are embedded, it is a precondition of propositional knowledge rather than identical to it. Yet since its failure makes propositional knowledge impossible, it is appropriate to describe a failure of Kantian recognition of certain particulars as ignorance of those particulars in Aristotle's sense. In this case, you could not be said to experience the gray blob at all.

If it is unclear whether any of the concepts constitutive of your current perspective would enable you to recognize this entity or not, then you can in fact neither identify this entity as of a kind with which you are already familiar, nor can you differentiate any such kind from it. The recalcitrance of this entity to identification in terms of the properties currently at your cognitive disposal calls into question all the concepts that form your current perspective: If they do not clearly fail to identify this entity, neither can they clearly succeed in identifying any other. So if you cannot now ascertain whether this entity is a street sculpture, three-martini lunch hallucination, or tropical plant, you cannot ascertain whether it is a gray blob or not, either; or whether, if it is not, anything else could be.

This conclusion may seem to be too strong. Surely, it might be objected, it does not follow from the fact that you do not know what something is that you therefore do not know what anything is. Indeed it does not. But the preceding narrative does not address the question of what or how you *know*, nor even what propositional *beliefs* you have; but rather a presupposition of both of those questions. It addresses the question of whether, if you cannot successfully *recognize* something you experience in terms of the concepts at your disposal, you can successfully recognize anything else you encounter at the same time; and concludes that the answer is no. If you cannot recognize something in terms of the concepts at your disposal, you cannot identify it as having the properties of which you have those concepts. In this case, propositional beliefs about, and *a fortiori* propositional knowledge of that thing are impossible.

Again it might be objected that it does not follow from the possibility that your identification of one thing is incorrect that your identification of everything else is called into question. Indeed it does not; yet again the objection misses the point. The preceding narrative does not address the question of whether your identification of something is *correct* or not – nor, therefore, the question of the fallibility of your other identifications; but rather a presupposition of both of those questions. It addresses the question of whether, if you fail to make something you experience *rationally intelligible* relative to everything else you experience at the same time, you can succeed in making anything else rationally intelligible at that time, despite this one failure. Again the preceding narrative concludes that the answer is no. If you cannot recognize something as the same as or different from something else, then you cannot identify that second thing relative to it. Hence the question of whether you have identified either one of them correctly or not does not arise.

But this may seem unsatisfactory. For it must be in some sense possible for us to recognize an unfamiliar thing in terms of its lower-order properties, independently of our ability to identify it in terms of higher-order ones; otherwise how could we ever come to recognize and

eventually categorize unfamiliar things at all? The implication would seem to be that if we were truly to adhere to the requirement of vertical consistency, we could never learn anything new. I defer addressing this valid objection to Section 7 below, after having developed a more fine-grained taxonomy of agent's perspectives than that which we now have. At that point I suggest that it is, indeed, much more difficult for some agents than for others to learn anything new, about themselves or anything else, even if, like the gray blob on West Broadway, it is staring them in the face.

That you do not experience at all what is rationally unintelligible to you is why your encounter with a gray blob on West Broadway will not necessarily plunge you into madness. The lesson of this encounter is, rather, one about the centrality of literal self-preservation: that we have a very deeply ingrained, motivationally effective aversion to rational unintelligibility, because it threatens the rational coherence of the self as having that experience. We saw in Chapter II that an agent must, as a matter of conceptual necessity, finally be able to conceive of everything that happens to him consciously as his experience, in order to conceive of himself as capable of altering what happens to him, and so in order to exercise his agency. The preceding narrative shows us that, conversely, an agent who cannot conceive of his experiences in a rationally intelligible form cannot conceive of them as his experience at all, and so lacks the agency necessary to change them. So having and exercising the self-consciousness property that makes the experiences conjointly constitutive of his perspective at a particular moment rationally intelligible is not only a necessary but also a sufficient condition of unified agency.

That the self-consciousness property is both necessary and sufficient for unified agency does not imply that it is sufficient for any particular action. As we have seen in Chapter V.4, a particular action is determined by the particular occurrent concepts and beliefs that antecede it. It does imply, however, that without the self-consciousness property, no particular action would be possible. An agent who experiences events that she cannot make rationally intelligible in terms of the concepts that constitute her perspective at a given moment *loses her perspective on those events*: She confuses them with others, and all of them with herself. That is, she confuses all of those events with the rationally intelligible cognitive and conative events that constitute her perspective at a given moment. But a self that confuses unintelligible external or internal events with itself loses the ability to distinguish those events from itself, and with it the ability to defend its rational integrity against them; and so, finally, the ability to act intentionally in response to them. That is why your most likely initial reaction to the gray blob on West Broadway will be neither madness nor annoyance, but instead temporary cognitive and conative paralysis. It is this cognitive and conative paralysis, and the loss of unified selfhood and agency it threatens, that motivate us either to render a perceived conceptual anomaly rationally intelligible at any cost – even at the cost of plausibility, accuracy, or truth; or else to suppress the perception altogether.

Thus in some ways literal self-preservation may seem to be an impossible task. We are continually assaulted, if not by the presence of gray blobs, by other internal and external anomalies that test the psychological strength of the self to withstand them, or its cognitive

flexibility to accommodate them within the constraints of rational intelligibility. And we must take it as a given that we can neither withstand all such events – on pain of the fate that frequently befalls ostriches who bury their heads in the sand; nor can we accommodate all of them – on pain of the fate that befalls overloaded computers, whose simulated cognitive psychoses bear a touching resemblance to our own. The necessary and sufficient connection between agency and theoretical reason thus confronts every such agent with the dilemma of her own imperfection: We cannot possibly make rationally intelligible everything that happens to us, nor everything we think, feel and do, without threatening the coherence of that which we think we do understand rationally. So we cannot possibly integrate all such events without undermining our agency. On the other hand, theoretical reason is all we have for coping with such cognitive assaults. The alternative would be passively to acquiesce in the threat of unintelligibility, disorientation, and ego-disintegration that such anomaly represents. Such an inclination towards literal self-destruction could have no survival value.

So the demands of theoretical reason must be attenuated and bent to the contours of our limitations. Its consistency requirements must remain in force, but be made easier to satisfy. The stringency of those requirements must be maintained, yet tempered by rational loopholes. Thus we systematically distort and truncate our intellects, with the help of our rational capacities themselves, so as to achieve the illusion of rationality. The result is not rational intelligibility in the sense described above, but rather pseudorationality. Pseudorationality is our only rational choice.

4. Denial and Theoretical Investment

From Chapter II.2 and the case of the gray blob just discussed, we have seen how denial might operate in cases in which the arsenal of concepts constitutive of an agent's perspective is completely inadequate to identify a conceptual anomaly. If one has no concepts even remotely appropriate for coping cognitively with the thing in question, one will simply fail to recognize that thing as an experience one has. Here the preservation of rational intelligibility, i.e. literal self-preservation, requires that one remain oblivious to its presence. So when I said above that in denial one fails to recognize the thing as an experience one has, I meant that one fails to identify the thing that violates the concepts constitutive of one’s perspective even as an instance of the concept, "violation of the concepts constitutive of my perspective." One certainly may have that concept, and apply it appropriately. But not in this case. For denial eradicates recognition of the anomalous object, event or state of affairs completely. Because the particular unavailable to conscious recognition rationally ought to be, I described denial above as *biased* nonrecognition, and its preservation of rational intelligibility as *degenerate*.

4.1. The Naïf

But now contrast the case of the gray blob on West Broadway with some in which denial is required, not in order to preserve the rational intelligibility merely of one's experience as such,

but rather the rational intelligibility of *a certain interpretation or theory of* one's experience. The distinction can be limned as follows. I may be able to make sense of everything I experience as my experience, just in case I can subsume each such lower-order concept of that experience under the highest-order concept of the self-consciousness property, i.e. of its being an experience I have. It is not impossible that I might do this without trying or being able to make sense of it as confirming the theory that, say, it's a jungle out there, or that everything happens for a reason, or that I am a serious person, or some more sophisticated theory of human nature, or the physical world, or myself. In the first case, the rational intelligibility of my experience is a function of its horizontal and vertical consistency relation to the highest-order concept of the self-consciousness property *simpliciter*: All the experiences I have are mutually consistent with one another relative to the concept of their being my experiences. This is the only highest-order concept that unifies all of them. I shall describe someone who conceives her experience in this way as a *naïf*.

The naïf lacks what I described in Volume I, Chapter VIII.3.2.2.2 as a *personal investment* in any particular theory of her experience. To review, an agent A is personally invested in something *t* if

(1) *t*'s existence is a source of personal pleasure, satisfaction, or security to A;
(2) *t*'s nonexistence elicits feelings of dejection, deprivation, or anxiety from A; and
(3) these feelings are to be explained by A's identification with *t*.

A identifies with t if A is disposed to identify *t* as personally meaningful or valuable to A. Thus having a personal investment in *t* implies, among other things, desiring *t*. As we saw in Volume I, Chapter II.2.1, desire is an incorrigible source of pseudorationality because it distorts perception of the desired object and everything around it, by magnifying the properties of the object that satisfy or frustrate the desire, and minimizing those which are irrelevant to doing so. And as we saw in Section 2 above, this is what naturally inclines human agents to evil on Kant's view.

To be personally invested in a theory implies desiring that the theory be true; and this, in turn, that the truth of the theory serve some further desired goal or end. Thus personal investment in a theory treats the truth of the theory as instrumental to the satisfaction of some further desire one has. Because the theory in question is a theory of one's experience, it implies that one's instrumental desire for the truth of the theory saturates all of one's experience, i.e. all the experiences that the theory is supposed to explain. Because all of one's experiences are saturated by an instrumental desire for the truth of one's beliefs about those experiences, one is exposed to all of the dangers and limitations of funnel vision, the permanent condition of the Humean self described in Volume I, Chapter II.2.3. Because this instrumental desire permeates the interpretation of all of one's experiences, it introduces cognitive distortion into all aspects of one's interiority, pulling one's perception and comprehension of everything off balance.

To say that the naïf lacks a personal investment in any particular theory of her experience is to say that there is no particular theory of her experience she desires instrumentally to be true; no particular theory she desires to justify, vindicate or rationalize her desires; and hence no

particular theory she desires to cater to them. This does not mean that the naïf has no desires. But it does mean that such desires play a negligible role in coloring her perception of reality – and, because most objects of desire are, in turn, determined by or constructed from perceptions of reality – therefore a negligible role in the motives that move her to action. Hence a naïf may approximate a fully effective intellect of the sort described in Chapter V.4.5. The naïf is freer to act on the motivational basis of occurrent thoughts or beliefs whose content determines and precipitates the content of her actions, such that those thoughts or beliefs are unconfined and unsystematized by the organizing and subsumptive functions of theory. Her primary intellective commitment is to reality itself, not any particular theory of it, nor to the satisfaction of any particular desire that that theory serves. So it is rather to reality itself that her actions respond.

For these reasons, the naïf is less prone to encounter genuine conceptual anomalies, and is not susceptible to theoretical anomalies at all. For the only requirement something must meet in order to be rationally intelligible as one of her experiences is that it must be the kind of thing she can, in fact, experience. Included among the lower-order concepts that constitute the naïf's perspective are the commonsense observational ones of size, shape, color, etc. we all share. But absent from that perspective are the kind of higher-order concepts that qualify and restrict the scope of those observational concepts to any particular theory of the kind of thing one can in fact experience. Because the naïf's perspective is not circumscribed by theoretically imposed restrictions on what can be seen, heard, felt, or done, the naïf does not need the mechanisms of pseudorationality to excise any of it. Since the naïf lacks higher-order pet theories that restrict what qualifies as, say, contemporary art, the possible mutant effects of radioactive fallout, a sentient being, or the latest advertising gimmick that a mottled gray, mooing blob on West Broadway might violate, she has less cause to suppress recognition of such a blob than you or I.

But then we already know, from folklore and history as well as from personal experience, that naïfs, and children, often see many things, not just the emperor's sartorial desolation, that the rest of us habitually overlook. In some non-Western philosophical traditions, the naïf's outlook is a sought-after and highly valued state of unmediated mental clarity and spiritual integrity that committed philosophers work long and hard to achieve. These traditions assert that to see things as the naïf sees them is to see them as they really are. To see such things as they really are is to grasp them directly and without preconceptions. To see and experience without preconceptions is to bypass the cognitive limitations on experience that make conceptual anomaly possible. Without any cognitively restricted system for filtering and organizing the data of experience at all, the very concept of an anomaly ceases to have application.

Of course the absence of such a system also implies, in the limiting case, the absence of all of the conditions for preserving the internal coherence of the self which we have already discussed. Thus to see as the naïf does may imply, in the limiting case, the temporary or permanent suspension, disintegration, death, or in any case absence of a unified ego. The canonical Western tradition of philosophical psychology contains no resources for explaining how an agent could have or survive such an experience, hence no resources for explaining how a

naïf as I have described it might be possible. Perhaps the closest analogue might be Kant's model of intuition, abstracted from its symbiotic relationship to synthesis; I discuss this model at length elsewhere. But even it is only an analogue. Other, non-Western traditions of philosophical psychology have a more sophisticated apparatus for explaining the experience of the naïf.[3]

Contrast the case of the naïf with a second, in which I do have a personal investment in some favored theory of my experience; henceforth I shall refer to this as *theoretical investment*.[4] The case of theoretical investment is different. For here the rational intelligibility of my experience is a function of its horizontal and vertical consistency over time relative to *two* higher-order concepts, the mutual relation of which may vary. First, there is the concept of things as experiences I have; and second, there is the concept of things as confirming my favored theory of my experience. What is the relation between these two? There are at least three possibilities, and so at least three kinds of theoretical investment.

4.2. The Ideologue

We have already seen in Chapter II.6 that the second concept could not dominate the first without violating a necessary condition of agency. Of course this does not mean it cannot dominate the first, period. By an *ideologue*, I shall mean someone who regards his experience as an instantiation of his theory, rather than the other way around. He thus has a sense of mystical inevitability about himself, as an impersonal force in the world that, like other such forces, behaves in the ways his theory predicts. The ideologue may seem to have the concept of the self-consciousness property, in that he recognizes things that happen to him as experiences he has. But in fact this recognition is hollow, because he does not, in so doing, recognize things as happening to him precisely in that form *in virtue of his nature*: Instead, he thinks his experience has the character it has in virtue of the forces, specified in the theory, that determine his nature. And he interprets his own active responses to that experience in similarly impersonal terms, not in terms of personal motivations to alter it. Because the ideologue accepts no responsibility for the particular character of his experience, in fact he does not fully grasp the concept of the self-consciousness property. Hence he abdicates a necessary condition of motivationally effective agency: His thoughts, feelings, and impulses are to him a series of *aha-Erlebnisse*, forced upon him by his situation; and he is, to varying degrees, propelled into action by impersonal forces that are beyond his interior control.

For the ideologue, theoretical anomaly is intolerable. By threatening the rational intelligibility of his favored theory of his experience, it threatens, so far as he is concerned, not only the rational intelligibility of that experience itself, but thereby the rational intelligibility of the universe and his predestined place in it. Because he regards his own experience as an instance of his theory, rather than the other way around, it is not open to him to rethink his perspective on the world as independent of that world itself. His perspective is such that he views it as fully determined by that world, in the ways specified by the theory that purportedly describes it. To undermine the theory, then, is to undermine everything at once. For the

ideologue, theoretical anomalies do not exist. I say more below about some more subtle pseudorational mechanisms by which they are made to disappear.

4.3. The True Skeptic

Like the ideologue, the character I shall describe as the *true skeptic* also attempts to make all her experiences rationally intelligible, relative both to her favored theory, and to the concept of the self-consciousness property. I describe this character as a *true* skeptic rather than merely as a skeptic, in order to distinguish her attitude toward theoretical anomaly from that derisive and dismissive one adopted under the guise of skepticism by the merely provincial. By contrast with the ideologue, the true skeptic reverses the relation between her favored theory and the concept of the self-consciousness property; for she recognizes her favored theory, and its confirmation by her experiences, as itself an experience she has. So even if her theory that, say, it's a jungle out there, or that she is a serious person does, in fact, make all of her experience rationally intelligible, she conceives it as doing so *in virtue of her nature*, i.e. as itself an experience she has: Her favored theory is subordinate to the highest-order concept of the self-consciousness property. Because of this order of priorities, the true skeptic's investment in any such theory can never be more than tentative, and her attitude toward it never more than pragmatic. If the theory makes sense of what is already rationally intelligible as her experience, well and good. If it is undermined by theoretical anomaly, then it is to be revised or replaced. But this is merely to restate what we already know about true skeptics, namely that on the one hand, they are, indeed, inclined to skepticism about higher-order explanatory theories; and on the other, fondly attached to the observational data those theories are recruited to explain.

Like the naïf, the true skeptic has less trouble with theoretical anomalies than the ideologue. For since she recognizes even her favored theory to have the character it does in virtue of her nature, her personal investment in it cannot be so absolute as to blind her to the possibility of its – and her – limitations. And since she lacks such an overriding personal investment in her favored theory, its modification or eventual replacement by a theory better able to accommodate the existence of gray blobs is more a matter of regret than anxiety or panic. Hence the true skeptic's attitude toward theoretical anomaly is one of curiosity and inquiry rather than suspicion or rejection. Finally, since her conceptualization of her experience as hers takes priority over her conceptualization of it in terms of any such tentatively held theory, she is, like the naïf, freer to recognize a gray blob simply for what it is.

4.4. The Dogmatist

The figure for whom the relation between the concepts of the self-consciousness property and of his favored theory as confirmed by his experience presents a genuine dilemma is one I shall call the *dogmatist*. Luckily, this dilemma is one the dogmatist is unusually well-equipped to solve. For the dogmatist, the relation between these two concepts is one of uneasy parity: Both are of the highest order in the dogmatist's perspective; neither is subordinate to the other. The

dogmatist both conceptualizes all of his experiences as his, and also conceptualizes them as instantiating his favored theory. The dogmatist would not deny that his experiences have the particular character they have in virtue of his nature. Nor would he deny that they have that character in virtue of the truth of his favored theory. Rather, the dogmatist would congratulate himself on the good fortune of being so constituted that the way he experiences the world is, in fact, the way it is. Thus for the dogmatist, these two concepts are materially equivalent.

The notion of being personally invested in one's favored theory about the world has special poignancy in the case of the dogmatist. For the dogmatist is someone who does derive very great pleasure, satisfaction, and security from his favored theory of his experience. Indeed, the dogmatist may feel instinctively that it is the only genuine source of security to be had. These feelings are, of course, to be explained by his identification with his favored theory. But notice that the higher-order priority he gives to his favored theory implies his identification with it in an even stronger sense than that required by the definition of personal investment. His favored theory of his experience is not just personally meaningful or valuable to him; it *is* him at the deepest level of self-identification. For as we have just seen, he assumes that the way he experiences the world is, in fact, the way his theory depicts it; and that this, in turn, is the way it is.

Theoretical anomalies that threaten or undermine the rational intelligibility of the dogmatist's favored theory are correspondingly anxiety-producing. For in so doing, they undermine the dogmatist's conception of his own experience, and the rational intelligibility of that experience itself. Thus the dogmatist is like the naïf and the ideologue (and unlike the true skeptic), in that all three are made more susceptible to rational disintegrity by their unqualified attachment to the concepts that constitute their perspectives. But the dogmatist is like the ideologue (but unlike the naïf and the true skeptic), in that the personal investment of both in favored theories of their experience constricts the scope of their perspectives, and so brings the threat of rational disintegrity that much closer. Because the favored theory with which the dogmatist strongly identifies restricts the range of concepts by which to make sense of those realities, his perspective on them is correspondingly less open-ended, more rigid, and therefore more fragile. The constriction and fragility of the dogmatist's perspective creates more occasions on which he may encounter theoretical anomalies, to the extent that his favored theory excludes more from its scope of rational intelligibility: modern art, ESP, the inscrutable cultural Other, avant-garde styles of self-presentation, play, astrology, jokes, games, interpersonal theater, agitprop cultural subversion, and his own delinquent impulses, must be either explained (or explained away) by his theory, or else consigned to conceptual oblivion. It is for the dogmatist, as for the ideologue, then, that the gray blob on West Broadway may present a real problem.

5. Denial as Biased Nonrecognition

We can now distinguish three circumstances in which denial may be an expected response to the presence of anomaly; only the last is, strictly speaking, a pseudorational response.

First, it may function as it does for the naïf, in order to exclude from consciousness something that is anomalous relative even to the most comprehensive and flexible concept one has, namely the concept of something as an experience one has. A conceptual anomaly that is not recognizable in these terms is by definition conceptually inaccessible, and so is not a possible candidate for rational intelligibility in the first place.

Second, denial may function as it does for the true skeptic who is in a loose sense a member of the scientific community, in that her favored theory of her experience has been tested, confirmed, and consensually validated to some extent by that community. Macroscopic determinism might exemplify such a theory. Theoretical anomaly relative to such a theory still may be an experience one has, but the weight of consensus and scientific method militate against acknowledging it as such. Under these circumstances, the anomaly may be a candidate for rational intelligibility, but the true skeptic's measured inquiry, plus the weight of collective theoretical reason itself is against it. Here again, denial is consistent with the requirements of theoretical reason.

Third, denial may function as it does for the dogmatist or the ideologue, whose theory may or may not receive consensual validation, but the biases of which in any case would not survive disinterested critical scrutiny. In this case we may, but need not, appeal to rational experimental method in order to determine this. Our commonsense, serviceable criterion for distinguishing that which is so obscure or genuinely enigmatic as to be rationally inaccessible from that which is intersubjectively obvious is *third-personal disinterested recognition*. If a third party, similarly equipped both culturally and cognitively, but lacking the dogmatist's personal investment in his favored theory, can make the thing rationally intelligible relative to her perspective, whereas the dogmatist cannot relative to his, then the dogmatist's difficulty is not that the thing in question is theoretically anomalous, but rather that his favored theory is just too restrictive or provincial to accommodate it. In this case, his denial of the thing in order to preserve the rational intelligibility of his theory is a pseudorational strategy, and the rational intelligibility thus preserved is degenerate.

Note that the test of third-person disinterested recognition tracks pseudorationality rather than provinciality. The favored theory that saturates a dogmatist's perspective and mediates his relation to his experience may be the most cosmopolitan one available. Yet his personal investment in it may cause him to fail that test when confronted by a particular phenomenon. Conversely, an agent's favored theory may be relatively provincial, thus multiplying the likelihood of encountering theoretical anomaly and therefore the opportunities for pseudorationality – yet survive that test in a particular case.

Because the dogmatist identifies his experience with his theory, rather than conceiving his experience as subordinate to it, he has further cognitive resources for meeting such challenges, in addition to pseudorational denial, that the ideologue lacks. We have already seen that because the ideologue lacks a necessary condition of agency, she lacks the conception of herself as actively *doing* things like thinking, inferring, and searching her memory. This is not to

say that she does not do these things at all; just that she does not conceive herself as doing them. Hence by contrast with the true skeptic, the ideologue does not conceive herself as capable of revising or rethinking her favored theory – or, by contrast with the dogmatist, as capable of rearranging it to fit the facts. Thus the ideologue is inclined to avoid incorrigibly anomalous experiences at all costs – through psychological self-insulation, skillful circumnavigation, or, when all else fails, simply shutting her eyes very tightly and magically thinking the anomaly away.

The dogmatist has the same cognitive resources for conceptually rearranging things as he had for arranging them in the first place, in order to satisfy the consistency requirements of rational intelligibility. And he is more highly motivated to do so, by the fragility and constriction of his theory, and his self-protectiveness toward it. That is, the dogmatist has not just a biologically fundamental disposition to render his experiences horizontally and vertically consistent over time, as the rational intelligibility of those experiences requires. In virtue of his personal investment in this favored theory, he has in addition a contingent but central *desire* to render his experiences horizontally and vertically consistent over time, relative to the requirements and constraints of his favored theory of those experiences. The more provincial his theory, the stronger this desire must be. Hence this analysis implies that the more provincial his theory, the more inflexible the preference for consistency that McClennen rightly dismissed as arbitrary.

For these reasons, the requirements of horizontal and vertical consistency over time afford the dogmatist the option of two more subtle pseudorational strategies, in addition to blanket denial, for dealing with theoretical anomalies. And his natural disposition to satisfy these requirements, plus his personal investment in his favored theory, motivate him to exercise those strategies. From now on, in discussing these two further pseudorational strategies, I speak not just about the dogmatist, but also about *us*. This is not because I think anyone who is likely to read this discussion is purely and simply a dogmatist in the sense described. Obviously, the naïf, the ideologue, the true skeptic and the dogmatist are all equally caricatures, abstracted from more complex agents whose dispositions, ends and perspectives may change from moment to moment, and who are capable of exhibiting the characteristics of each. But I do think that anyone likely to read this discussion probably does have a favored theory of her experience, however nascent or inchoate; a theory in which she is, to varying degrees, personally invested. So I hope to be analyzing cognitive phenomena that all of us will recognize.

6. Dissociation as Biased Negation

Our disposition to satisfy the requirement of horizontal consistency supplies us with the pseudorational strategy I call *dissociation*. Recall that horizontal consistency requires us to conceive all our experience at a given moment as mutually logically consistent, i.e. as satisfying the law of noncontradiction. Relative to a favored theory of that experience, this is to require, first, that the theory be horizontally consistent; and second, that all our experience be

recognizable in the theory's terms – i.e. that they be vertically consistent. A theoretical anomaly is then by definition anything that defies recognition in these terms. This is one juncture that separates the dogmatist from the true skeptic. The true skeptic's tentative investment in his theory allows his greater detachment from it, in order more easily to rethink or revise it in order to accommodate what appears to be a theoretical anomaly. By contrast, the dogmatist's personal investment and self-identification with her theory makes her reluctant to abdicate or modify it, and inclines her to construe her theory, and therefore the events and phenomena it explains, honorifically, as normative goods. Relative to these, the negation of her theory a theoretical anomaly represents is to be dismissed not only as an intrinsically alien, inscrutable enigma, but as therefore *insignificant*, without value, and so unworthy of further attention. The familiar slide from a phenomenon's inexplicability to its unimportance, so well documented by sociologists of science like Kuhn, should disabuse us of the conviction that there is any hard and fast distinction between the descriptive and the normative, at least in practice, to be made.

For example, art viewers and critics are notably susceptible to dissociation in the judgments they make about works of art. Dissociative judgments about art tend to have a backhanded valorizing function for the works excluded from them. Here the concept of art is taken to be interconnected with other, specifically normative concepts such as quality, value, or beauty; so that to identify an object as art is thereby to ascribe to it an honorific status. Conversely, to remark about an object, "That's not art," is not merely to make a value-neutral observation about a curator's taxonomical error. It is to disparage the work on the grounds that it lacks those normative properties of quality, beauty, etc. that would valorize it as art. Past and present victims of this type of dissociative judgment include folk art, fiber art, women's art, African American art, Native American art, Indian art, Aboriginal art, African art, primitive art, ethnic art, crafts, happenings, outsider art, conceptual art, photography, film, video art, digital art, multimedia art, performance art, body art, street art, graffiti art, environmental art, earth art, and so on. These terms often function as convenient labels for expressing one and the same dissociative concept, of non-art. They thereby dismiss the work as a legitimate candidate for aesthetic appraisal. As a rule, such dissociative judgments relegate to the status of non-art those classes of objects that do not already enjoy considerable institutional legitimation. However, art viewers are hardly alone in deploying dissociation as a cognitive tool for narrowing the scope of judgment to that which is institutionally authorized. Claims that memoirs are not real literature, or that compositional minimalism is not music, or that women or colored people or gays or Jews or Arabs are not legitimate candidates for various social roles or professional positions do the same.

In all of these cases, dissociation plays the same role, of culling from one's scope of judgment conceptually disruptive particulars – whether objects, ideas, or people – and relegating them to the cognitive sidelines, where they are disregarded. The resulting, blinkered view of reality is one of disintegrity because cognitively sidelining an object, idea or person divides and destabilizes our awareness. The more crowded the cognitive sidelines become, the more they

encroach on the priorities on which we are trying to maintain exclusive conscious focus. Agency itself is undermined by dissociation, because the objects, ideas, or people we mean to dissociate are of course objects of our awareness. By dissociating them, we dissociate part of our own awareness from itself.

Thus dissociation maintains some minimal degree of rational intelligibility for the event, object or state of affairs, but nevertheless fails fully to integrate it into one's theory-laden perspective. In this case the thing is not lost to consciousness altogether, as it is in denial. But it is disconnected from the agent's own conception of the range of objects of his experience. It is conceptualized as the negation, and so the devaluation, of some concept or set of concepts that defines the agent's theory-laden perspective, rather than positively in terms of those concepts. Although the agent can positively predicate the self-consciousness property of a dissociated event, he cannot positively predicate other significant properties of it. So he can say of a dissociated event that it is an object of his experience. But he cannot say substantively and positively what kind of experience it is – only what kind of experience it is not. The concepts necessary for making it rationally intelligible are only minimally and negatively available to him. His perspective is not broad enough fully to integrate it. This is why the horizontal consistency within the positive theory that dissociation succeeds in preserving is degenerate.

This account of dissociation is compatible with Philip Bromberg's. Bromberg says of dissociation that "the experience that is causing the incompatible perception and emotions is 'unhooked' from the cognitive processing system and remains raw data that is cognitively unsymbolized within that particular self-other representation..."[5] I go further than Bromberg in claiming that although the "raw data" in question is cognitively unsymbolized within what he also calls a "unitary self-experience" – what I would call an agent's theory-laden perspective at a particular moment, this does not mean that it cannot be conceptualized at all. Dissociation is distinct from denial in that dissociated anomalies, but not denied ones, can be conceptualized; but only in negation-concepts that disconnect that object from the rest of the agent's conception of her own perspective.

These observations apply in social situations as well as in personal or interpersonal ones. Dinah, for instance, is, on the one hand, acknowledged as a member of her social community – whose social behavior at dinner parties, on the other, conforms to different conventions. After dinner, everyone else moves the furniture out of the way and dances, while Dinah stands on the sidelines, watching and making witty conversation. When she is invited to join in the dancing, she declines. From the perspective of other agents in that community, Dinah's behavior is conceptualized simply as a violation of established customs: as obstructive, inhibiting, or rejecting of those customs. It is not that these other agents do not recognize her behavior at all. It is simply that they can make it rationally intelligible only in terms of concepts of what it is not: It does not facilitate sociability, it does not encourage informality or vulnerability, it does not promote mutual participation, it does not satisfy everyone else's desire to participate in a shared and inclusive group experience, it does not aid digestion or burn calories, and so on. In

application specifically to the circumstances at hand, the theory-laden social conventions of this community do not include the positive concepts of resting (rather than exercising) after a big meal, or of deepening social contact through verbal (rather than ritualistic physical) exertion, or of detached observation as a valid social role. So Dinah's behavior is conceptualized only negatively, and therefore disapprovingly, in terms of the theory-laden conventions and social purposes it violates. Her community lacks the occurrent concepts for conceiving it positively, in terms of those it might, under other circumstances, promote.

The problem here is not that the positive concepts of resting after a big meal, etc. are not available anywhere in the social environment. Indeed, agents in this community might make use of them themselves under different circumstances. The problem is rather that these positive concepts are not occurrently available to their perspectives for making Dinah's behavior in these circumstances rationally intelligible. The situation is in this respect like any in which we regard another person's behavior as incomprehensible because our perspectives lack the modal imagination and so the impartiality to conceive of what it might be like to be in her shoes, even though in fact we often are. The resulting provinciality of the community's theory-laden perspective on Dinah's behavior leads them to conceptualize her lack of participation in the shared social ritual of after-dinner dancing merely as *contradicting* the concepts that define the social theory in which they are invested. Hence it violates the horizontal consistency of that theory. It is because others conceive Dinah's behavior as *inconsistent* with their conceptualization of the situation – that dancing after dinner is appropriate and beneficial – that they experience it as a threat, criticism or offense to their social unity, rather than neutrally; and because they experience it as a threat that their conceptualization of it counts as biased negation.

This is always true of dissociated anomaly, whether it falls at the normal or the pathological end of the spectrum. Similar considerations would apply to pathological dissociation of the sort found in, for example, multiple personality disorder. Clearly there is a wide range of cases. But all seem to involve either the denial of behavior that is completely unintelligible relative to the agent's self-conception; or else the dissociation of behavior that is intelligible only when conceived as being performed by some other "self" that is almost always conceived as subsidiary to and inconsistent with the primary "self." Successful psychotherapy then reintegrates the separate perspectives of each "self" into one more comprehensive one, and eliminates those that are strictly incompatible with it.

Reconsider, finally, the gray blob on West Broadway. There are, obviously, a variety of ways of making sense of this entity, and we have considered some of them. But it is equally easy to construct a rather arid theory of one's experience in which there is simply no room for such things: a theory, say, in which there are two sexes, three races, a circumscribed set of acceptable roles and relations among them, an equally circumscribed set of acceptable norms of behavior, dress, and creative expression, and a further division of the human race into those who observe these standards and those who do not. Not only gray blobs, but much else that is of interest, not just in our contemporary subcultures, but in other ones as well, will then fall outside the pale of

this theory. Again, someone with a personal investment in such a theory similarly will tend to dissociate such phenomena from the realm of the meaningful and important, and consign them instead to the status of intrinsic and uninteresting conceptual enigma – assuming that these perceived enigmas do not allow their existence to be denied altogether.

7. Rationalization as Biased Predication

I described *rationalization* as a degenerate form of vertical consistency. Recall that vertical consistency requires us to preserve transitivity from the lower-order concepts by which we identify something to the higher-order ones they imply. Relative to a favored theory of our experience, this is to require, first, that the lower- and higher-order concepts of the theory be vertically consistent, and second, that any experience recognizable in terms of its lower-order concepts instantiate the relevant higher-order ones as well. Now any theory even ostensibly worth its salt must include, among its lower-order concepts, the observational concepts by which we commonsensically interpret our experience: of shape, color, size, and so forth, however otherwise provincial that theory may be. But this means that even a provincial theory of one's experience can exclude only genuine conceptual anomalies, of the kind that might trouble the naïf or the true skeptic, through its lower-order concepts. It cannot exclude gray blobs simply by fiat.

This may explain the valid objection, noted in Section 2 but not addressed there, to the case of the gray blob on West Broadway as originally narrated: Surely, we felt, if we have the lower-order concepts of grayness, shapelessness, mooing things, and so forth, we can recognize the thing in question as a gray blob, even if we cannot say what higher-order kind of gray blob it is. Indeed, provincial theories were characterized as precisely those that made into conceptual anomalies things that were well within the range of rational intelligibility from a theoretically disinterested perspective. The need for rationalization arises because the commonsense rational intelligibility of these things at lower conceptual orders puts pressure on the theory's higher-order concepts to accommodate them, on pain of violating the requirement of vertical consistency, and so of revealing the conceptual inadequacy of the theory. The dilemma for one who is personally and dogmatically invested in such a theory is that she must accommodate the anomaly without seeming to revise the higher-order concepts of her favored theory; this dilemma is what separates the dogmatist from the true skeptic. It is for the dogmatist that rationalization is of greatest use: It is the process by which one stretches, distorts, or truncates the customary scope of instantiation of the higher-order concepts of one's theory, in order to accommodate the recalcitrant phenomenon within the theory's scope of rational intelligibility.

Consider, for example, Jensen's and Murray's theories of the putatively inferior intelligence of African Americans. Now that we know the very concept of race itself to be without foundation in genetics, merely a pseudorational fiction developed in order to rationalize seventeenth-century slavery in the Americas, Jensen's and Murray's theories look even dimmer. Based on the relatively low mean scores of African Americans on standardized intelligence tests,

they both dismiss the overwhelming evidence of environmental influence in favor of the now fully discredited notion of genetically inherited racial characteristics. However, it is generally agreed that such tests incorporate a cultural bias along many dimensions that limit their diagnostic use to the assessment of competence at performing only certain very specific and rather circumscribed tasks, namely those required by the tests themselves. The resulting clinical concept of "intelligence" is used only with scare quotes by most legitimate diagnosticians.

In order to derive from performance on such tests the conclusion that African Americans are of inferior intelligence in the broader, socially honorific sense, one must redraw that honorific concept of intelligence very narrowly. Jensen's and Murray's concept of intelligence must exclude, for example, the ability to not only survive but in many instances flourish in a lethally hostile social environment in which one is outnumbered ten or more to one; to formulate and carry out complex, ambitious, and self-interested long-term plans of action under highly adverse conditions deliberately designed to thwart them; the ability not only to grasp but successfully to use, analyze, and refine abstract cultural concepts that may be completely alien to one's native or familial environment; to learn, adapt, and creatively develop predominant but unfamiliar social and cultural practices to one's own benefit as well as to that of the predominant alien culture; and so forth.

These are the true tests of intelligence, and would that we were all so successful in passing them. To redraw the concept of intelligence so narrowly that it excludes such abilities is to save the coherence of the theory by sacrificing the plausibility of the concepts that compose it. But it is not difficult to spot the background, provincial theory of which Jensen's and Murray's are crudely rationalized refinements, nor the theoretical anomaly those rationalizations are designed to accommodate. The background theory is of the arid sort described earlier, in which acceptable roles, relations, and behavior between African Americans and European Americans are conceived in such a way that the very idea of a resourceful, creative, insightful, flexible, ambitious, highly competent African American is by definition theoretically anomalous. The theory accommodates the anomaly by redefining the scope of the honorific concept of intelligence so as to exclude it.

Or consider once more the gray blob on West Broadway. Again it is easy to imagine a theory of a particularly self-righteous and sour-minded sort, according to which this blob is, like so much else on West Broadway, nothing but one more capitalist plot to poison the minds of the unsuspecting masses and fill the coffers of media devils. The beauty of any favored theory of one's experience is a boon for the personal investor in provincial ones, namely the versatility of its constituent concepts. Pseudorationality, if not genuine rationality, is an available resource for literal self-preservation for even the most dogmatic and narrow-minded among us. For as Humpty Dumpty knew, we are free to use concepts in any way we like.

8. Pseudorationality in Application

The following chapters are concerned primarily with the operation of pseudorationality as a response to first-person theoretical anomaly, i.e. to the self-protective measures we take against first-personal violations of our moral theory, and so against violations we ourselves commit against our morally inflected self-conceptions. I address both the violations themselves and, in Chapter X, the adequacy of the moral theories thus violated. So it will be convenient to both close this chapter and preview the following discussion by examining how all three of the pseudorational mechanisms just enumerated operate in tandem under such circumstances.

Here is a real life example, as described by John Maynard Keynes. The scene is the Paris Peace Conference of 1919, in which the Allied powers – Great Britain, the United States, Italy, and France – are deciding how to carve up Germany and what amount of reparations for World War I are to be demanded of it. The question at issue is whether Germany should be required to reimburse the Allies for the pensions and separation allowances they pay to widows of soldiers who died in the war. On the face of it, this is an unusual and unwarranted request, since a country that goes to war may be presumed ordinarily to be responsible for shouldering the financial benefits it promises its soldiers as a condition of their enlistment. And Germany is already being forced to pay a great deal more than is consistent with jump-starting its economy sufficiently in order to make those reparation payments in the first place.

The French under Clemenceau, the Prime Minister, are clearly committed to extracting enough in reparations to permanently cripple Germany's economic system and reduce its population, so there is no irrationality in their insistence on this policy. But the Americans, who have the least to lose, are, under the leadership of President Woodrow Wilson's previously published Fourteen Points, committed above all else to doing what is just and right according to that document. President Wilson then allows himself to be persuaded by Clemenceau that Allied expenditures on pensions and separation allowances count as war damages inflicted by Germany on Allied civilian populations, hence should be included in German reparation payments.

The Germans justifiably protest that this is inconsistent with the prior terms of assurance implied by the Fourteen Points, and on the basis of which they formally surrendered. "But this," Keynes comments,

> was exactly what the President could not admit; in the sweat of solitary contemplation and with prayers to God he had done *nothing* that was not just and right; for the President to admit that the German reply had force in it was to destroy his self-respect and to disrupt the inner equipoise of his soul; and every instinct of his stubborn nature rose in self-protection. … It was a subject intolerable to discuss, and every subconscious instinct plotted to defeat its further exploration.
>
> Thus it was that Clemenceau brought to success, what had seemed to be, a few months before, the extraordinary and impossible proposal that the Germans should not be heard. If only the President had not been so conscientious, if only he had not concealed from himself what he had been doing, even at the last moment he was in a

> position to have recovered lost ground and to have achieved some very considerable successes. But the President was set. … it was harder to de-bamboozle this old Presbyterian than it had been to bamboozle him; for the former involved his belief in and respect for himself.[6]

First, the mechanisms themselves. *Rationalization*: Wilson allowed himself to be convinced that having to pay one's own soldiers' pensions was a war damage that Germany had inflicted on Allied civilians, a case of doublethink funhouse reasoning if there ever was one. *Dissociation*: Wilson responded to the Germans' warranted protest against this bit of bad-faith sophistry by, in effect, closing down the psychological borders, by isolating and recasting his own failure of rational autonomy as "*nothing* that was not just and right," relative to the rational intelligibility of his morally inflected self-conception. *Denial*: And finally, in order to maintain his personal investment in his theory of his own virtue, Wilson simply denied the Germans the opportunity to dilate upon this reproach by denying them the chance to speak out against it at all, by refusing even to entertain further discussion along these lines.

Second, the in-tandem operation of these mechanisms. When one's dogmatic investment in one's theory of oneself is very deep, i.e. when that theory serves and satisfies very deep desires for moral rectitude and the satisfactions that a sense of moral rectitude brings, the threefold arsenal of pseudorationality offers a powerful resource for defending this theory against the transgressive incursions of one's own moral imperfection. Whereas rationalization stretches and twists the terms of the theory out of recognition in order to cover and thus validate the delinquent behavior, dissociation aids this by negating contradictory characterizations of it that would threaten this reinterpretation; and denial, backed by force and political authority if necessary, eliminates them from consideration. Conjointly these three mechanisms serve to protect and preserve the theory against the pressure of doubt, re-evaluation, self-interrogation and revision, when the psychological and political price of such destabilizing self-criticism would be too high. However, I think Keynes in this narrative is uncharitable to Wilson in suggesting that it was merely the latter's self-respect and belief in himself he felt compelled to protect. The proclamation and publication of Wilson's Fourteen Points had raised exaggerated international expectations of him and of American participation in the Paris Peace Treaty negotiations that would have been impossible to fulfill under the best of circumstances. The cost of raising such expectations and then dashing them is not simply a loss of self-respect and self-confidence, but rather the crushingly humiliating knowledge that one has failed one's fellow man, disappointed expectations that one knowingly encouraged them to have. Thus pseudorationality defends one against self-recognition of the fumbling moral arrogance that invites dishonesty with others. It is this basic self-deception that lies at the foundation of pseudorationality to which I turn next.

Endnotes to Chapter VII

[1] The term is Kant's, but he makes a variety of uses of it. See 1C, B 368 (conceptus ratiocinantes), A 339/B 397 (vernünftelnde Schlüsse), A 406 (vernünftelnde Schlüsse), B 449 (vernünftelnde Lehrsätze), B 450 (vernünftelnde Behauptungen) A 462/B 490 (vernünftelnde Behauptungen), A 497/B 526 (vernünftelnde Argumente); and G Ak. 405 (wider Gesetze der Pflicht zu vernünfteln).

[2] Compare Kuhn's discussion of scientists' responses to anomaly in scientific theories in Thomas Kuhn, *The Structure of Scientific Revolutions* (Chicago: The University of Chicago Press, 1970), especially 62 – 66, 78.

[3] I find that detailed in the *Yoga Sutras* to be the most authoritative and convincing. See any of the translations and commentaries listed in the Bibliography for references. Prabhavananda and Isherwood's (New York: Mentor, 1969) is at once the most accessible and inspiring for a Judeo-Christian audience and also, with regard to translation, probably the most misleading.

[4] Chapter V.5 implies that it does not matter whether my favored theory of my experience is normative or explanatory; I address this question more fully in Chapter X, below. But briefly, any powerful explanatory theory also prescribes a way things are supposed to, i.e. *should* work under ideal conditions, and so contain a normative component. And any full-blooded normative theory also explains a way things *would* work if conditions were, in fact, ideal, and so contains an explanatory component. Part of what we do by attempting to make things rationally intelligible in the terms given by our favored theory of our experience is to assess the extent to which the real measures up to the ideal – or, to put it in Hegel's infamous terms, the extent to which the actual is rational.

[5]"'Speak up that I may see you: Some reflections on dissociation, reality and psychoanalytic listening," *Psychoanalytic Dialogues* 4 (1994), 517-547; pp. 520-521.

[6] *The Economic Consequences of the Peace* (Mineola, New York: Dover Publications, 2004; orig. London: Macmillan and Co., 1920), 49-50.

Chapter VIII. First-Person Anomaly

So far I have focused on pseudorational responses to third-person conceptual anomalies. These attempt to restore the rational intelligibility and coherence of our perspectives on the world against the threat posed by conceptual anomalies in the external environment, by tinkering with the contours of the favored theories that mediate those perspectives. Our pseudorational responses to first-person anomalies are more complex, because the entity trying to restore rational intelligibility and coherence to the favored theory is identical to the entity violating them. First-person anomaly violates the theoretical rationality of that part of a favored theory that explains sentient, animal, specifically human behavior of the even more specific sort that the agent conceives herself to instantiate – with respect to gender, ethnicity, physical type, character, personality, social stratum, occupation, and so on. In short, first-person anomaly violates the agent's interior, morally inflected self-conception. This interior self-conception is the agent's favored theory of herself; and attitudes or behavior that remain rationally unintelligible in its terms are theoretically anomalous relative to it. Our own theoretically anomalous attitudes, emotions and behavior pose a more immediate – or better, a more entirely unmediated – threat to our self-conception as unified agents than do enigmatic external events. Of course this does not mean that they qualify as true conceptual anomalies in the sense defined in Chapter VII.4.1.

From Chapter VII.4.1 – 4 we have seen that not all agents necessarily have a personal investment in their favored theories. Hence not all agents necessarily have a personal investment in their self-conceptions. I further explore this kind of case in Section 1.1, below. But we also saw in Chapter VII that the mechanisms of pseudorationality are prompted only when an agent does have such an investment; and are most poignantly and complexly prompted when that investment in his favored theory is a dogmatic one in the sense defined in Chapter VII.4.4. So the same considerations mentioned about dogmatists in general apply with special force in the case of dogmatic responses to first-person conceptual anomaly. This case is the focus of the following discussion. First-person conceptual anomaly relative to an agent's favored theory of himself does not necessary render the enterprise of literal self-preservation self-subverting. But it does create unintelligibility in the agent's self-conception. The threat of interior disintegrity therefore ramifies much farther in the first-person case: not only between one's theory-laden perspective and veridical perception, but also between the cognitive mop-up operations of pseudorationality and the internal cognitive, conative and affective anomalies that require them.

I remarked in Chapter VII.3 that pseudorationality was our only rational choice in the face of continual external assaults – of conceptually anomalous events and information – on the rational integrity and coherence of the self. That was not strictly true. In theory, it is open to us to abdicate our personal investment in our favored theories sufficiently simply to endure the anxiety, confusion, disorientation, and powerlessness that often accompany reminders of our subjective fallibility. That is, it is psychologically possible simply to abdicate the aspiration either to inviolable agency, or to infallibility, or to unalloyed moral rectitude. The naïf, and to a lesser

extent the true skeptic show us how literal self-preservation and so rational intelligibility thereby *might* be vindicated in the end. But it is not possible abstractly to assign relative probabilities to the consequences of either letting go of these aspirations, or stubbornly digging in.

Reminders of our subjective fallibility are much harder to endure, if being right is more important to us than being genuinely rational – if we are, indeed, dogmatists. The stakes are even higher if the theory about which we need to be right is our theory about ourselves; if we console ourselves too often with the thought that although we may not be perfect, we at least know whom we are. If self-knowledge, i.e. being right in one's self-conception, is even more important to one than being right about other things, then the lure of pseudorationality will be all the more compelling. The more importance we accord to such self-knowledge, the more susceptible we are to pseudorational judgments about what our obligations are, and whether we have fulfilled them. A strong personal investment in any aspect of our self-conception, assaulted and undermined by enigmatic or personally unacceptable attitudes, beliefs, emotions, or actions, will call forth an even more intensified mobilization of the resources of pseudorationality to withstand it. This is the phenomenon we understand as *self-deception*. Briefly, self-deception is our pseudorational response to first-person theoretical anomaly.

Section 1 offers an analysis of self-deception as pseudorational belief about first-person anomaly, contrasts it with the standard analysis, and applies it to an extended fictional example. Section 2 extends the analysis from pseudorationality about conceptually anomalous belief to pseudorationality about conceptually anomalous emotion, motivation, and action; and proposes a solution to the problem of moral paralysis raised in Volume I, Chapter VIII.2.2. Section 3 describes pseudorational reaction to morally anomalous action in the first-person and the third-person case, grounds an account of one (among many) origins of evil on the perverse asymmetry between these two cases, and contrasts this account with Nietzsche's. Section 4 selectively reviews Aristotle's, Kant's and Nietzsche's accounts of pseudorationality, with particular attention to Kant's. This part of the analysis builds on the analysis of moral impartiality in Chapter VI as requiring symmetrical interiority between the first- and third-person cases; and explores several ways in which the symmetry requirement may be violated and the agent's interior integrity consequently destroyed. Section 5 adopts the perspective of the anomalous self that is the object of these pseudorational operations, and Section 6 concludes that rational integrity is not an in-theory impossibility in the non-ideal case. Finally, Section 7 applies this conclusion, building on the apparatus developed in Chapters II and III, to show how our disposition to preserve rational integrity both imposes constraints on rational final ends and so terminates the infinite regress of self-evaluation discussed in Volume I, Chapter VII that is generated by Frankfurt's Humean conception of the self.

1. Self-Deception

A personal investment in our self-conception is a personal investment in its horizontal and vertical consistency over time. This investment requires that, at any given moment, we *conceive* the experienced things and properties our self-conception subsumes in such a way as to satisfy the requirements of theoretical reason, *whether they do so in fact or not*. This, in turn, strongly disinclines us to detect logical inconsistencies in our theory-laden conceptions of our experience. In particular, this investment renders us unable to conceive ourselves at a particular moment as simultaneously desiring contradictory objects, nor as simultaneously believing contradictory propositions, even if in fact we do. If this is true, it means that for dogmatists with a personal investment in their self-conceptions, self-deception is just as inevitable as self-consciousness. Below I explain the sense in which self-deception is pseudorationality about first-person theoretical anomaly. For in situations in which we may simultaneously hold such contradictory beliefs or desires, it is virtually impossible for us to recognize this.

Self-deception is a particularly difficult and central problem for metaethics because, as Keynes showed us in Chapter VII.8, no matter how fully developed or compelling our substantive moral theory may be, it is useless to us if we are psychologically incapable of acknowledging that we have violated it. "A conscience," Alice Hamilton observed, "may be a terrible thing in a man who has no humility, who can never say, 'I might be mistaken.'" Kant also saw this quite clearly. He saw that the really pressing motivational problem for actual moral agents – i.e. for motivationally ineffective intellects – is not weakness of will, but rather self-deception.[1] Kant realistically assumes weakness of will to be a given – just as I do motivationally ineffective intellect in the non-ideal case. For Kant, the question is how to deal constructively with this given. Kant concludes, pessimistically, that we cannot deal with it constructively at all; I discuss his account of the pseudorational mechanisms of self-deception in Sections 4.1 – 3 below. My account builds on Kant's. Both regard weakness of will as a metaphysical given, and both regard self-knowledge about weakness of will – i.e. knowledge of our moral derelictions with respect to our actual motives and obligations – as at the very least cognitively unusual. In my account, the rarity of self-knowledge about weakness of will is rooted in my analysis of self-deception as an antecedent, pseudorational cognitive dysfunction that obstructs it.

1.1. Selfless Dogmatism vs. Self-Deception

Above I identified dogmatists as especially susceptible to self-deception. However, not all dogmatists are self-deceivers, because not all dogmatists are personally invested in their self-conceptions. Consider a cult member. A cult member self-identifies with a dogmatic and provincial theory of her experience; a theory in which her degree of personal investment necessitates denial, dissociation, or rationalization of dissonant data, in order to preserve the rational intelligibility of her experience. She might also have a self-conception with which her favored theory is interdependent. Nevertheless, such an individual might be *selfless*, in the sense that her pseudorationality is motivated solely by her dogmatic allegiance to the theory, and not

by considerations of personal vanity or self-esteem. She might, indeed, simultaneously exhibit all the beneficent virtues to a particularly high degree: devotion to others, sympathy, generosity, humility, modesty, and so forth; virtues that lead us to deplore all the more their being squandered in the service of the dogmatic theory that deludes her.

To call the cult member selfless is not to say she lacks a self, for it is precisely the virtuous characteristics of the self she expresses whose waste we deplore. Nor is it to say that she lacks a self-conception, for she conceives herself as, among other things, devoted to the dogmatic and provincial theory that commands her cult membership. Rather, it is to say that her self-identification with her favored theory is not itself motivated by self-aggrandizing considerations. While she defends her self by pseudorationally defending her theory, the defense of her theory is not intended to redound to her own greater glory. Conversely, although an assault on her theory is an assault on the rational coherence of her self, she does not perceive such an assault as a personal *insult*, nor as denigrating her own value. Her responses to such an assault include anxiety and panic, not rancor or resentment. That the cult member's personal investment in her pseudorational theory is to be explained by her selfless self-identification with it, but not her self-aggrandizement by it, underwrites the intuition that this case is, indeed, most naturally described as a case of delusion, not self-deception. To identify it as a case of self-deception would be conceptually peculiar.

The implications are two. First, although all self-deceivers are dogmatic pseudorationalizers, not all dogmatic pseudorationalizers are self-deceivers. The cult member has everything it takes to be a dogmatic pseudorationalizer, but lacks a certain feature conceptually necessary to being identified as a self-deceiver. Second, therefore, self-deceivers are dogmatic pseudorationalizers of a certain kind: They are dogmatic pseudorationalizers with a personal investment in a certain kind of dogmatic theory, namely one with two mutually dependent parts: The first, explicit part is a dogmatic and provincial theory of their experience, of the sort already discussed. The second part, often left implicit, is their self-conception: the theory of who they are, how they behave, and how they relate socially to others. A self-conception in which an agent is personally invested therefore contains incorrigibly honorific and self-aggrandizing (if not self-congratulatory) components. It is the self-deceiver's personal investment in this second part of the theory, a self-conception that is mutually interdependent with the provincial theory of his experience, that is the source of the vanity and false pride the cult member was shown to lack.

This second part of the theory is not to be confused with the self-consciousness property. The latter is merely the value-neutral concept of one's self as having one's experiences; the former is a substantive, honorific conception of the kind of self one is; for example, that one is a serious person, or is fair-minded and tolerant in one's judgments, or does only what is just and right. Any agent may have a self-conception, and not all self-conceptions function as does the dogmatist's. A dogmatic self-conception, the unstated second part of the self-deceiver's theory, is mutually dependent with the first, in that the validity of the first is a necessary and sufficient

condition, in the self-deceiver's eyes, of the validity of the second. This is because, typically, the first part, the dogmatic theory of his experience, includes in it honorific status for persons of the kind he conceives himself to be.

On this analysis, then, a self-deceiver is a dogmatic pseudorationalizer who conceives of himself as a good and valuable person if and only if the dogmatic theory of his experience that he espouses is the correct one. Nazis, racists, misogynists, anti-Semites, and other elitists of various kinds are all obvious examples of individuals we might identify as (at the very least) self-deceived according to these criteria. But there are many other dogmatic theories of one's experience that may function similarly to thus align one on the side of the angels, as it were, depending on one's social values. It may be that, held by the right agent, any such theory may, in that agent's eyes, confer on him the exalted status of being holier than thou.[2] So the self-deceiver is that particularly beleaguered brand of dogmatic pseudorationalizer for whom the mechanisms of pseudorationality must suffice to preserve not only the rational integrity, but also, therefore, the honorific status of his self-conception.

1.2. The Standard Analysis of Self-Deception

Now one implication of the foregoing characterization of self-deception as a species of pseudorationality is that a certain familiar analysis of self-deception, as the case in which one believes that not-P because one wants to, even though one knows in some sense that P, is inadequate to the psychological facts. Either we must continually vacillate between believing that P and believing that not-P, adjusting our current perspective, favored theory of our experience, and self-conception accordingly, in order to preserve horizontal and vertical consistency; or else our personal investment in believing that not-P leads us pseudorationally to deny, dissociate, or rationalize P, in order to maintain the belief that not-P. In that case, I would argue, it is not true that we also "in some sense" believe or know that P. For to have any such belief would presuppose the rational intelligibility of P that our pseudorational mechanisms are designed to obscure.

The second implication of the foregoing analysis is that, even if we could be said to "in some sense" believe or know that P while believing not-P because we want to, as the standard analysis would have it, this analysis could not in any case provide a sufficient condition of self-deception. For according to this standard analysis, we would have to identify the cult member as self-deceived, which, as I have suggested, seems conceptually peculiar. In addition, one's desire to believe the falsehood not-P must be, specifically, a desire for self-aggrandizement, to which belief in the falsehood is a means. This is to argue that in addition to deception of the self by the self, self-deception also intrinsically involves deception *about* the self that deceives.[3]

Is there any pseudorationality recognizable as self-deception that does not involve self-aggrandizement? I doubt it, but remain open to persuasive counterexamples. Consider two kinds of case, nonpersonal and personal. First the nonpersonal case: Suppose I have a personal investment in the theory that it's a jungle out there. Also suppose, for the sake of argument, that

this theory is false. My investment in it then may be explained, either by the particular, generally oppressive experiences I and most everyone else seem to be having; or by the fact that this theory excuses my own failures and moral derelictions. In the first case I am merely mistaken in my beliefs. Only in the second case does it make sense to describe me as self-deceived.

Now take the personal case. Suppose I have a personal investment in the theory that my spouse is a good person. Again suppose this theory to be false. Again my investment in it may be explained in at least one of two ways: either by my spouse's resourcefulness in maintaining an appearance of virtue and guilelessness, which elicits my love and respect, or by the fact that my recognition of his moral turpitude would reflect negatively on my conception of my own tastes, preferences, and susceptibility to moral corruption. If my spouse is recognizably a bad person, then either I have vicious tastes – say, a fascination with evil, or else the close proximity of evil leaves me morally unconcerned. In this case I have self-defensive and self-aggrandizing motives for deflecting any such recognition. Again it seems appropriate only in this case to describe me as self-deceived.[4] Hence self-deception does not depend on the content of the theory in which one has a personal investment, but rather on the motive that causes the investment. My thesis is that it always involves the desire to buttress another, usually unspoken theory, namely an honorific personal self-conception. The de facto rational consistency of experience alone is not enough for the self-deceiver.

1.3. Test Case #2: *The Margin*

Next I consider a fictional example that is identifiable as one of self-deception according to these criteria, and test the capacity of the foregoing analysis to explain it. Take the hero of Andre Pieyre de Mandiargues's *The Margin*.[5] Sigismond, while on a business trip in Barcelona, has received an ominous letter from his wife Sergine's servant. As he begins to open the letter, his eyes alight on the sentences, "She ran to the wind tower. She climbed the spiral staircase. She threw herself from the top. She died right away." He decides not to read the letter just yet, and puts it in a prominent place on his hotel dresser. For the next three days, he drifts through the streets of Barcelona, revelling in its museums, architecture, and unsavory nightlife. Some of his experiences recall to him with disgust his dead father's depravities. Often he finds himself imagining Sergine's sturdily impassive reactions to the situations he encounters, responding as he imagines she would, and reminiscing fondly about episodes in their life together. Every morning he returns to his hotel room, naps, notices the letter, and goes out again. Sometimes he thinks about the letter there in his hotel room while engaged in very different pursuits. His revelry is gradually brought to a halt as his companion of the night deserts him, his pleasures grow stale, and the image of the unopened letter becomes more persistent. Finally he returns to the hotel, and opens and reads the letter, to learn that his only child Elie has drowned in an accident, and that Sergine, immediately upon discovering this, has committed suicide. He quits his hotel, drives away from Barcelona, and pulls over to the side of the highway, where he, too, commits suicide by shooting himself in the heart.

Now on the standard analysis of self-deception, we would be forced to describe Sigismond's state during his three days of revelry and dissipation as one in which he in some sense knew that Sergine had committed suicide, but convinced himself that she had not, because he loved her and did not want her to abandon him, and so believed both that she had (perhaps unconsciously) and that she had not. But this just seems completely inadequate to handle the complexity of the case. Sigismond may not have wanted Sergine to commit suicide, but surely this desire would ordinarily motivate him to ascertain whether she had or not, and, if so, why. And if he believed she had, why did he spend three days partying in Barcelona before committing suicide himself?

A different analysis is in order. First, the functioning of the pseudorational mechanisms themselves: The sanguinity of Sigismond's perspective is violated by the intimation of tragic news about his wife, in the form of the letter. He pseudorationally *denies* this intimation, with the help of the distractions and novelties his stay in Barcelona provides. Relative to the fragile and studied innocence of his perspective, he regards the physical presence of the letter on his hotel dresser as a potential threat that he pseudorationally *dissociates* as an inscrutable, enigmatic object that regularly intrudes on his disingenuity, only to be repeatedly dismissed. The exhaustion of his resources for denial forces him to confront the contents of the letter, in the hope of integrating it into the sanguine perspective he has, with the aid of these pseudorational mechanisms, so tenuously maintained. This proves to be impossible. Sigismond's avoidance of the contents of the letter is not predicated on his unconscious knowledge of its contents, but rather on his cognitive inability to make its contents rationally intelligible, relative to the constraints of his self-conception. These contents are threatening to him, not because he already knows what they are, but because he cannot find the conceptual resources for figuring out what they are, without violating the dogmatic assumptions in which he is personally invested.

Second, the personal investment that motivates Sigismond's pseudorationality: It is very hard to understand the point of Sigismond's pseudorational behavior without knowing the self-conception its presence threatened. After all, he cares deeply about Sergine; why wouldn't he hasten to find out whether the phrases in the letter actually referred to her, and, if so, what had motivated her suicide? The implication is that it could not have been news of Sergine's suicide alone that he was avoiding. It is similarly difficult to understand why the contents of the letter lead him to commit suicide himself, without reference to his self-conception. After all, his affection for his son Elie was rather distant to begin with; and although Sergine's suicide must be a terrible blow, he obviously is not without resources for containing his loneliness. The implication is that it was not just the combination of his wife's and his son's deaths itself that led him to this end. Without reference to the self-conception in which Sigismond is personally invested, we cannot quite understand why he has been so energetically motivated to deceive himself in the first place.

The description of the case provides evidence for what this self-conception is. We know, for example, that he feels both attracted and repelled by the thought of his own father, and that

he does not give a thought to his own son's safety after receiving the letter. We also know that he is, on the one hand, deeply attached to his wife; and on the other, untroubled by occasional, casual betrayals of her. Although his recollections of her include no demonstrative expressions of her love or affection for him, we know that he assumes that she is attached to him as well, and ignorant or tolerant of these dalliances. We can say, then, that he has a deep personal investment in the conception of himself as Sergine's beloved; of their bond as intimate, loving, and durable; that he views his extramarital activities as unproblematic, and is untroubled by Sergine's likely reactions to them. We also know that he feels some distaste for, or at least detachment from the role of father, and is emotionally indifferent towards his son.

That this self-conception is pseudorational is suggested, first, by the distance and impassivity of Sergine's responses as Sigismond has recalled them. They do not provide evidence of her emotional attachment to him at all. His assumption that she does love him is sustained by *rationalization*, by misconceiving her imperviousness as itself the way she expresses her love for him. This rationalization enables him falsely to assume that she loves him, because she does not correct it by telling him explicitly that she does not.

Second, the pseudorationality of Sigismond's self-conception is evinced by Sergine's having committed suicide immediately upon Elie's death – i.e. with not a moment's hesitation or thought for Sigismond's wellbeing. The implication is clear that without her son, Sergine's life is no longer worth living; and her husband, despite his attentions to her, does not make it so. Sergine's suicide nullifies by a single act the importance of his commitment to her as he conceived it, and thereby his value and importance in his own eyes. It is not simply the combination of her suicide and his son's death that drives Sigismond to suicide, but the now-inescapable realization that he meant so little to her that his love provided her with no consolation or further reason to live. In demonstrating through her suicide that he provides *her* with no reason to live, Sergine has taken away *his* reason to live. Sigismond is goaded to suicide by the realization that his self-conception as the valued and beloved object of her devotion was false; that in fact he is of value to no one whose opinion matters. This is the truth that he went to such lengths to avoid; that Sergine's suicide makes inescapable; and that makes his own suicide inescapable as well.

What makes Sigismond a self-deceiver, then, is not just that he manages to avoid unpleasant truths because he prefers not to know them, as the familiar analysis would have it. What makes him a self-deceiver is his self-aggrandizing self-conception, sustained by denial, dissociation, and rationalization: by a studied obliviousness to the conclusive, tragic evidence of his wife's indifference; by dissociation of the letter that contains it; and by rationalization of the earlier unresponsiveness to him that otherwise would have indicated it. His personal investment in his pseudorational self-conception is self-deceptive because it enables him to avoid recognition of who he really is.

1.4. Self-Deception and Self-Knowledge

But why is it in general so important for the self-deceiver to avoid self-knowledge? My thesis explains this by the self-deceiver's personal investment in a self-aggrandizing self-conception, in conjunction with the disparity between that self-conception and what the pseudorationalized evidence in fact indicates is a less exalted truth. In Chapter V.2.2 I argued that our highest-order disposition to literal self-preservation made the horizontal and vertical consistency of our favored theory of our experience tantamount to a normative good; and in Chapter VII that this disposed us to ascribe to it, and to the things it explains, an honorific status. I also argued there that a particularly fragile or provincial theory elicits an even more intensely self-protective desire to preserve it, proportional to one's personal investment in it. For these reasons, the self-deceiver is particularly recalcitrant and impervious to any attempts of her own to survey and critically revise her own pseudorational self-conception. Her investment in it is too great, and increases not only with its fragility, but also with the bogus value it confers on her. This is why the project of convincing a self-deceiver that she is self-deceived often seems such an exasperating and futile one: The self-deceiver has not only the rational intelligibility of her experience, but her self-conception as a valuable person, to protect.

But the same vigilance and self-protectiveness that leads the self-deceiver so strenuously to avoid self-knowledge leads her to value it all the more. For of course her pseudorational self-conception would become a source of intense humiliation to her, if it were revealed to be false: The revelation that one is not as nice, smart, or popular as one thought is a shaming experience, in which one's deficiencies are exposed to the ridicule of the cruelest and most unsympathetic spectator of all. To avoid this revelation, one must be either very humble on principle, like Uriah Heep, very vigilant, like St. Augustine, or, like the self-deceiver, very resourceful in one's commitment to truth. As Sigismond's case suggests, self-deception, and pseudorationality more generally, requires energy, perseverance, an inquiring mind, a good grasp of the data, and a deep desire for epistemic rectitude. In order to avoid the humiliation of self-discovery, the self-deceiver needs not only to excise the damaging evidence that portends it; but also to believe that the pseudorational mechanisms by which she does so themselves rather bespeak her honesty, sincerity, and perspicacity. Her pseudorational self-conception, then, provides not only a source of bogus value for the self-deceiver, but also the illusion of a limited but impregnable scope of personal infallibility that enhances it. Thus may self-reflection and a commitment to truth supply a pseudorational disguise for the self-deceiver. This is what I meant when I suggested, at the beginning of this chapter, that the self-deceiver would rather be right than rational.

Now against such self-deception, as well as other forms of pseudorationality, philosophers of a Humean persuasion, such as Sidgwick, Rawls, Brandt, and of course, Hume himself[6] have urged a palliative, i.e. vivid reflection on the relevant data in a calm and composed setting. But if the mechanisms of pseudorationality function as I have suggested, the Humean palliative may in many cases amount to little more than ineffectual bootstrap-pulling. For the whole point of exercising our pseudorational resources is to restrict what counts as relevant data

to the psychologically and theoretically palatable. If the self-deceiver, and the pseudorational agent more generally, had appropriate conceptual access to these data in the first place, vivid reflection on them would be unnecessary. For the self-deceiver, vivid reflection on the relevant data is an occasion for pseudorationality, not an antidote to it.

2. Affective and Conative Anomaly

So far I have discussed the pseudorational response to theoretically anomalous beliefs about oneself relative to a self-aggrandizing but internally coherent self-conception. Belief is the primary case because, as we have seen in Volume I, Chapter II as well as in Chapters II – VI of this volume, the Kantian conception of the self I am defending in this project claims that all aspects of an agent's perspective are mediated by the attempt to preserve the rational intelligibility of the concepts – and so the judgments, and so the beliefs – constitutive of it. But I tried to show in Chapter II that beliefs are complex intentional attitudes composed of subsentential constituent concepts; and, in Chapter VII and Section 1 above, that the horizontal and vertical consistency of these concepts can be violated before propositional beliefs or judgments are ever formed. So anomalous beliefs are only one type of conceptual anomaly that may conflict with an agent's self-aggrandizing self-conception and thereby call forth the self-deceptive mechanisms of pseudorationality. Emotions (including desires) and actions also must be represented conceptually in order to achieve rational intelligibility. So conceptually anomalous emotions and actions may have the same disruptive effect. Self-preservation requires the internal coherence and intelligibility of all of our experience, whether cognitive, affective or conative.

2.1. Affective Anomaly

We saw in Chapter VII that the cognitive principle of understanding external events causally is horizontally consistent with that of understanding interior events causally, by seeking out their origins in our upbringing, social environment, and previous experiences. We saw also that it is vertically consistent relative to the more general principle that subsumes both external and interior causal inquiry, i.e. that we understand all the phenomena of experience by seeking out their causal connections. But now consider how a theoretically anomalous interior event such as an unacceptable emotion might violate that part of our self-aggrandizing self-conception that describes our emotional character and so lead similarly to self-deception. It is a truism that we are socially and biologically disposed to delight in the esteem or admiration of a person we love; and similarly disposed to feel self-confidence and optimism upon receiving praise from some superior whose authority we respect. The more general, motivationally effective principle of which both of these are vertically consistent instances hypothesizes a positive, joyful response to obtaining approval from someone whose regard is valuable to us. We take this principle for granted in our self-conceptions, despite the reality that we do not always respond emotionally in the way we recognize as appropriate.

Suppose a highly valued personage in one's life shares too many extrinsic traits in common with other individuals one has valued highly in the past who have responded negatively to one's quest for approval. Suppose, for example, that he resembles one's wicked stepfather, hated sibling, or parasitic former spouse. Then one may respond to his esteem or praise, sought-after and highly valued as it clearly is, not with delight or self-confidence, but instead with rage, resentment, or the suspicion of ridicule. One's intuition that such emotions are inappropriate to their immediate causes may then lead one to *deny* or suppress them, or to refuse to identify them for what they really are. Thus one may express one's resentment in the form of sarcasm or verbal abuse, and claim, upon being confronted, that one was only joking, meant no harm, that one's victim is oversensitive or insecure, and so on. Alternately, one may *rationalize* one's anger by calling attention to the person's irritating imperfections, and claiming, for example, that anyone who speaks in a high whine and wears chartreuse is bound to provoke blind fury, no matter what his virtues. Finally, one may simply *dissociate* or disown one's inappropriate emotional response, by claim that it overtook one as a blind, irresistible impulse, and was completely outside one's ability to control. Self-deceivers who take this last tack tend not to recognize the inconsistency involved in then promising that it will never happen again.

Or consider the self-avowed "close friend" who sells one's confidences to the tabloids for a hefty fee, then purports surprise at the suggestion that her governing emotion toward one might be competitive envy, resentment, vindictiveness or greed rather than friendship – denying, perhaps, that selling one's confidences to the tabloids is inconsistent with strong emotional attachment of a benevolent nature; or dissociating her opportunism as an uncharacteristic moment of emotional immaturity; or rationalizing it as a well-intended promotional effort on one's own behalf. In both cases, the questionable behavior uncovers emotions at odds with the agent's self-aggrandizing theory of herself as mature, tolerant, and secure in her self-esteem. In both cases, the agent's own emotional response is theoretically anomalous relative to a self-conception that includes commonplace assumptions about psychological normalcy and socially appropriate behavior; but that may not be similarly anomalous relative to a more informed, sophisticated or self-reflective theory of human nature. An agent who is not too incapacitated by her personal investment in her favored theory of herself to do the hard work of analysis and introspection might well move from the relatively provincial, "pre-shrunk" self-conception in which these responses have no place to a more inclusive and informed self-conception in which they are to be acknowledged and controlled rather than disowned. The inclination to self-deception would be correspondingly diminished.

2.2. Conative Anomaly

Similar considerations apply to anomalies in action relative to one's conception of one's own character dispositions. Suppose, for example, that I conceive myself as a fair, generous, and sympathetic individual, and that most of my actions square with this morally inflected self-conception: I am in fact loyal to my friends, actively concerned to promote others' well-being,

and so on. However, I also spread unfounded and damaging gossip about individuals I dislike, thereby causing them severe personal and professional distress. This behavior would seem to be a clear instantiation of a motivationally effective principle that is horizontally inconsistent with those governing the rest of my conduct, and so violates the morally inflected self-conception they define. My highest-order disposition to literal self-preservation may lead me to defend the internal coherence of my self-conception by denying, perhaps sincerely, that I behaved in this way at all; or recall the behavior but deny that it is an instance of spreading unfounded or damaging gossip. Rather, I may rationalize it as merely an instance of indulging confidentially in harmless speculation – thereby denying as well the very real damaging consequences of that behavior, and ultimately my own responsibility for bringing them about. Or I may rationalize my conduct by arguing, say, that everyone gossips without thereby victimizing their subjects, and that no one need worry who has nothing to hide (thus defending the implicit thesis that anyone who is damaged by unfounded gossip must have deserved it). Finally, I may dissociate my behavior from that constellation of motivationally effective principles and concepts I identify as my self. By pleading that I am neurotic and easily threatened by others, and that mobilizing a network of social condemnation against them is a self-defensive reflex over which I have no control, I locate the cause of my behavior outside the scope of my voluntary agency.[7] In this case, too, my delinquent behavior is theoretically anomalous relative to a favored theory of who I am at which the cynical or misanthropic might snort. A more inclusive theory that rendered my gossip-mongering fully intelligible might be more informed or cosmopolitan, but not necessarily any more forgiving for that. A diminished inclination to self-deception brings with it a heightened taste for unflinching self-appraisal.

These self-defensive mechanisms for resolving internal incoherencies are just as inadequate to integrate first-person affective and conative anomalies in an agent's socially circumscribed or morally inflected self-conception as they were to integrate first-person anomalies of belief; and just as inadequate to integrate third-person anomalies in our theory-laden perspective on the physical world. All put a strain on the self that forces it to engage in yet more elaborate and irrational attempts to preserve its coherence – as, for example, when I conclude from the phenomena of quantum physics that all events must be random and all regularities illusory; or when I attempt to cultivate an attitude of emotional indifference towards anyone whose approval I in fact value highly; or when I offer for sale to the tabloids *all* of my friends' confidences, in order to demonstrate the moral innocence of having made a killing on yours; or when I ascribe to the person I have maligned through gossip a malevolent power to make me feel guilty. These self-deceptive responses to the internal incoherence of the self are irrational because they themselves ramify that incoherence yet more widely throughout the structure of the self, and motivate yet more elaborate attempts to ameliorate it; attempts that are similarly doomed to failure. Such pseudorational tactics can become so pervasive and overpowering that they can swallow up the self whose tactics they were – thereby replacing the unified subject whose perspective was overridingly governed by the highest-order self-

consciousness property with a tangled and incoherent mass of pseudorational defenses no longer capable of weighing from a distanced perspective their psychological costs. The threat of ego disintegrity thus generates a stance of vigilant, self-protective defensiveness that fails in direct proportion to its extent. The more incoherent and pseudorational the behavior of the self, the more vulnerable to such threats it becomes.

2.3. Behavioral Anomaly and Moral Paralysis

For an imperfect but motivationally effective intellect, acknowledging the delinquent behavior of the self as irrational is the best strategy for preserving the self against radical disunity, for this is to recognize that behavior as the painful threat it is to rational intelligibility. Because of the primacy of the highest-order disposition to literal self-preservation, a dawning recognition that the unity of the self is being destroyed by its own behavior disposes it, over the long term, to modify that behavior accordingly. In actual fact, it is questionable whether we ever truly succeed in reforming our conduct, without the prodding of these painful insights into our own irrationality. Those whose prior pseudorationality is so extensive as to render themselves incapable of such a recognition are correspondingly beyond the reach of self-reform.

Thus not all actual selves are free to exploit the option of self-reform. Although I argued in Chapters II.3 and V.4.4 that all selves are disposed to satisfy the cognitive requirement of rational intelligibility, it does not follow from this that all selves are disposed to satisfy the linguistic rule prescribing correct *use* of the *concept* of rational intelligibility. Hence not all actual selves may be disposed to conceive themselves as overridingly committed to rational intelligibility *per se* (even though in fact they are), nor to apply that concept correctly to their own behavior. Then the existence of demonstrably irrational behavior may not suffice to ensure its rational modification. Perhaps one may believe, rather, that being a sensitive or virtuous individual, or being interesting, or politically committed, is more important than anything else. And then one will feel impelled, under attack, to defend one's behavior at all costs in these terms, even in the face of glaring inconsistencies, and regardless of the psychological discomfort it causes one to do so. One will be disposed to deny, dissociate, or rationalize any evidence that undermines this defense.

Of course this self-deceptive response itself will strongly indicate that those values did not, in fact, have primacy in one's hierarchy after all. For in this case, the defense of one's own behavior requires the suppression or distortion of one's values in the service of pseudorationality, and thus sacrifices them for the appearance of rationality. But it is precisely the appearance of rationality that the self-deceiver is, on my thesis, most centrally disposed to preserve. Any such principles that are not vertically and horizontally consistent with the principles of theoretical rationality will be sacrificed similarly, in order to preserve the appearance of rationality against the reality of the self's interior disintegrity. This is the point at which the inadequacy of the utility-maximizing model of rationality by itself becomes very clear: Enormous sacrifices in all of the nonvacuous indices of utility – time, money, energy, reputation,

human resources, credibility – may be sacrificed in order to maintain the illusion of rational coherence. One (but far from the only) case study in the public sector would be the dedicated corporate and employee behavior of the Philip Morris tobacco company in the 1990s.[8]

Thus do we resolve in practice the problem of moral paralysis raised in Volume I, Chapter VIII.2.2. In fact, we are seldom torn by conflicting dispositions of the self, or inhibited from acting by uncertainty about our moral rectitude. More frequently, self-deception simultaneously resolves the conflict and ensures our moral rectitude by appealing to some conceptualization of our actions that succeeds in preserving their coherence with the rest of our behavior, and thereby permits us to keep peace with our consciences. It is only to the extent that we fail recognizably to preserve coherence that we are led, by our highest-order disposition to literal self-preservation, to change our ways.

3. Third-Person Moral Anomaly and an Origin of Evil

The above examples describe situations in which an agent's own unethical behavior is theoretically anomalous relative to his morally inflected self-conception. This is the more common case of moral anomaly, in which our self-conception is overly generous in giving us the benefit of the doubt, and thereby provides pseudorational cover for doubtful behavior from which we benefit. However, a self-aggrandizing self-conception that is thus morally inflected significantly impedes the ability to evaluate external moral action performed by another that really is conceptually anomalous – i.e. anomalous not relative to one's own provincial self-conception, but instead relative to the criterion of third-person disinterested recognition proposed in Chapter VII.5 for identifying genuine conceptual enigmas. I refer, of course, to the genuine conceptual anomaly of truly altruistic moral action, in which the agent behaves virtuously not only when it is convenient or costs her nothing; but when she does so despite the fact that the costs are significant. The whistle-blower cases discussed in Volume I, Chapter VI.5.2 and in this volume's Chapter VI.8 above would be paradigmatic examples.

In Volume I, Chapter II.2.4 I argued that the Humean psychology of desire conditions the agent to perceive the jousting tournament of desire-satisfaction as a zero-sum game, in which my gain is your loss and vice versa. And in Chapter VII.4.1 above I argued that unlike literal self-preservation, which is the object of a highest-order disposition, a self-aggrandizing self-conception is the object of a desire – more specifically, the object of a futile desire to avoid the self-hatred I argued is endemic to the Humean self. Hence preservation of the rational coherence of a self-aggrandizing self-conception requires not only pseudorationality, but also, thereby, explicit devaluation of external moral anomaly that threatens it. Devaluation of whistle-blowers and other such moral anomalies is expressed in (among other things) disparagement, demonization, ostracism, rejection, ridicule, retaliation, and physical violence. The magnitude of devaluation is directly proportional to the threat to self-aggrandizement this moral anomaly represents. The magnitude of threat to self-aggrandizement is, in turn, inversely proportional to the success of the pseudorational mechanisms that attempt to buttress it – by denying, for

example, the moral wrong the whistle-blower attempts to publicize; or dissociating him and his actions from the scope of one's legitimate moral community; or rationalizing them by ascribing to him self-interested or opportunistic motives for whistle-blowing: greed, revenge, public attention, and career advancement being the usual suspects. Of course these pseudorational mechanisms also have more harmful, overt behavioral analogues: of denial in cover-ups; of dissociation in ostracism or excommunication or shunning; and of rationalization in doctored evidence or character assassination or attacks on the whistle-blower's credibility or legitimacy. These operations of pseudorationality are directed primarily at the whistle-blower himself; similar ones also may be directed at the moral wrong he attempts to right.

The self-deceiver is thus betrayed by moral anomaly from both sides: from her own interior, by the first-person theoretical anomaly of delinquent behavior that contradicts her honorific but provincial self-conception; and from the external environment, by the third-person conceptual anomaly of actual virtue that makes a mockery of it. Her self-conception is undermined by her own unethical behavior as well as by others' ethical behavior, and by the further unethical behavior she herself undertakes in order to suppress them. And her moral judgment is correspondingly perverted by her need to cook up ways of commending and valorizing the first while condemning and devaluing the second. It is in this way, by finally arriving at the point at which the self-deceiver feels compelled to praise vice and condemn virtue in the service of self-aggrandizement, that the entirely innocent disposition to preserve the rational intelligibility of experience disposes the self-deceived dogmatist not only to become evil, but also actively to promote it.

On my Kantian account, then, evil is a consequence of moral self-deception, and moral self-deception is a consequence of the highest-order disposition to literal self-preservation in the non-ideal case, in which horizontal and vertical consistency are subverted by first-person moral anomaly. Evil is the expression of pseudorational defense of one's favored theory of oneself against the external, third-person conceptual anomaly of ethical behavior that, by contrast, threatens to make salient the unethical nature of one's own. The expression consists in pseudorational devaluation of the third-person case that satisfies the desire for self-aggrandizement in the first-person case. Because this expression of evil is itself a theoretical anomaly relative to that morally inflected self-conception, it, too, must be honorifically pseudorationalized through the infliction of yet further damage on the third-person case. Thus the asymmetry between the virtue of the third-person anomaly and the vice of the first-person anomaly is compounded by the agent's increasingly energetic, and so futile attempts to pseudorationalize the third-person anomaly out of existence. Think of pseudorationality in this context as a quicksand that sinks the agent in ever deeper in a morass of vicious behavior, the more she grasps at psychologically and behaviorally repressive pseudorational mechanisms to extricate herself from it.

On Nietzsche's account, by contrast, evil is a consequence of the same conditions that engender interiority, namely self-control in response to external oppression. Self-control requires

the interiorized agent to internalize rather than spontaneously to express anger and resentment against the external, spontaneous agent that oppresses him. These internalized emotions ripen into hatred and the desire for revenge against her. They then are sublimated into a demonized representation of the oppressively spontaneous other onto which the interiorized agent projects his murderous rage, by ascribing to the demonized other fantasy motives of deliberate malice that purport to explain her oppressive behavior. However, the interiorized agent's proto-hypothesis about the oppressively spontaneous other is fundamentally mistaken, on Nietzsche's account: in fact, spontaneous agents possess neither the intellectual abilities nor the psychological complexity necessary to be deliberately malicious or oppressive. The worst they can be is careless, insensitive, negligent or stupid. The fantasy motives of deliberate malice that the interiorized agent projects onto oppressively spontaneous others in fact describe only his own vindictive impulses to retaliation. Representing the other as evil incarnate leads him to become evil incarnate himself.

Because rational interiority is a species of interiority more generally, it may seem at first glance that my account of the origin of evil is a species of Nietzsche's. However, the very possibility of self-control that engenders interiority presupposes the success, to some degree, of literal self-preservation in the non-ideal case – and therefore the rational intelligibility of the agent's experience at subsentential levels. So in fact Nietzsche's account presupposes mine. Moral self-deception is a necessary pseudorational condition for the demonized fantasy projection of the other that legitimates the interiorized agent's vindictive hatred of her, and the repressive pseudorational mechanisms he exercises in order to gratify it.

4. Kant (and Others) on First-Person Moral Anomaly

My Kantian account of the origin of evil presupposes Kant's own analysis of the morally pernicious effects of pseudorational self-deception. I believe this analysis, in turn, is presupposed by Kant's account of evil in *Religion within the Limits of Reason* alone, but I do not defend that belief here. Instead I focus on Kant's rather unsystematic account of the three pseudorational mechanisms themselves, as they operate on first-person moral anomaly. In all three cases, pseudorationality serves to justify violation of the symmetry requirement on interiority that, as we saw in Chapter VI, essentially defines impartial moral principle.

4.1. Kant on Rationalization

Here is Kant's account of how rationalization in reaction to first-person moral anomaly bespeaks interior disintegrity. He observes in the *Groundwork* that

> [i]t is indeed at times the case that after the sharpest self-examination we find nothing that without the moral basis of duty could have been powerful enough to move us to this or that good action and to so great a sacrifice; but from this it cannot be inferred with certainty that it is not some secret impetus of self-love which has actually, under the mere false pretence of this Idea, been the actual determining cause of our will. We gladly

> flatter ourselves with a falsely accommodating nobler motive, but in fact we can never, even by the most strenuous examination, fully get behind our secret drives; …[G, Ak. 407]

Despite the suspicion that our beneficent act may have been motivated by unacceptably self-interested or self-aggrandizing considerations, Kant says, we convince ourselves that our action was motivated by ethical principle rather than personal politics.

Let us take an example. We have a moral obligation to respect the uniqueness and singularity of each individual we encounter. We have an obligation to recognize them as who they are and treat them accordingly: to not confuse our life partner with our father or mother, not treat salesclerks or other service providers as though they were inanimate instruments of our will, not view friends and colleagues merely as service providers. This obligation is entailed by Kant's fourth formulation of the categorical imperative, that we are to treat others' humanity as an end in itself [G, Ak. 429].

We may sincerely wish to live in a world in which everyone's uniqueness is respected. We may deeply believe that human beings should not be treated as though they were prefabricated items on an assembly line. These convictions naturally assume special salience when we ourselves are so treated; when we feel insufficiently acknowledged or valued for the particular combination of needs, goals, talents, and idiosyncrasies that define us. We may experience this failure of basic regard in a wide variety of circumstances. At the trivial end of the spectrum, there is the customer service representative who cuts off our question before we have finished asking it with an irrelevant formulaic answer that fails to address it. At the serious end is the friend who rewards us with approval when we satisfy his legitimate needs, but resists the reciprocal obligation to satisfy ours. Both kinds of mistreatment and all of those in between are painful, but not only because they devalue us morally. They are painful because they fail on a more elemental, epistemological level to see us clearly, and so fail even to meet the fundamental requirements for genuine intersubjective communication. The resulting feeling, of interacting with oneself alone in a vacuum, is extremely unpleasant.

Yet we often treat others in this way, applying convenient preconceptions or behavioral formulas that obscure another's singularity. We may not interrupt our interlocutor with a formulaic answer. But we may still call it forth at the appropriate pause in the conversation. We may not overtly resist our reciprocal obligation to satisfy another's legitimate needs. But we may still chafe under it silently, or find ways to subvert or evade it. In such cases we implicitly judge respect for the other's uniqueness to be inconvenient, inefficient, or inconsistent with the promotion of our own best interests. Despite the frequency with which we may make such morally partial judgments, we generally are not eager to acknowledge that we do; and indeed may evince genuine surprise, or even heartfelt outrage, at the mere suggestion. When such judgments nevertheless intrude on the rational intelligibility of our theory-laden perspectives on ourselves, we find a way to explain them away.

In such a case, there are many "nobler motives" with which we may flatter ourselves. In dealing dismissively with a salesclerk, for example, we may tell ourselves that we are merely respecting the boundaries of privacy and impersonality between two strangers who wish only to perform a business transaction as quickly and efficiently as possible. Under the "mere false pretence of the Idea" of respect for her privacy, we may arrogate permission to run roughshod over her essential singularity. After all, this is what the telemarketer does to us, when he repeats by rote the same sales pitch to anyone irrational enough to answer the telephone before the machine picks up. We can be sure such "nobler motives" are a bit of self-aggrandizing flattery because there is of course no necessary conflict between respecting another person's uniqueness, and respecting her boundaries of privacy. But our ennobling self-conception may relegate this obvious fact to the status of theoretical anomaly.

For Kant, what goes wrong in such cases is that we misapply the concept of respect for the impersonality of a transaction to what is in fact a violation of the moral obligation to respect others' singularity. We do this through biased predication, by distorting the scope of the concept of respect for privacy or impersonality. We magnify the properties of the situation that instantiate this concept – for example, making much of the fact that our interaction is with an anonymous salesclerk who surely has no interest in forging a deeply authentic connection with us. And we minimize those properties that fail to do so – for example, that she is elderly, has been on her feet behind a counter all day; probably works without commission for $6.50/hour, and so on.

Philosophers may be particularly susceptible to the temptations to rationalize away such first-person anomalies of behavior, because of the intellectual agility we learn as part of our training in reasoning and analysis. We like to play at being Humpty Dumpty, making words mean what we want them to mean, revising those definitions when they no longer serve our purposes, and formulating and reformulating moral principles accordingly. Philosophers may be more nimble than lay people in these precarious intellectual activities, because of their training. But we are hardly alone. Philosophy as a discipline is merely a rational reconstruction of informal theorizing about what goes on in other parts of the self, in addition to what goes on in the world outside it. The use of euphemism and spin is hardly a specifically philosophical vice, and by now it should be clear that rationalization is not different in kind from euphemism or spin.

So our training as philosophers does not enable us any more easily to resolve the interior conflicts and anxieties that are expressed in rationalization, whether inflicted on ourselves or on our audience. These anxieties are exacerbated by the violence we do to concepts and principles by using them to rationalize theoretically anomalous ethical violations in biased defense of our high opinion of ourselves. For they signal to us that the conventional association between term and referent, concept and particular, is being broken; and so that the needs of understanding and communication are being subordinated to the requirements of self-defense. These signals ramify the internal disjunctions between principles and practice into a much larger disjunction between

mind and world. That is why we often say about self-deceivers who are truly adept at rationalization that they are "out of touch with reality," or "living in their heads."

4.2. Kant on Dissociation

Kant's discussion in the *Groundwork* also describes the interior rational disintegrity that biased negation can wreak. He says,

> If we now attend to ourselves whenever we transgress a duty, we find that we in fact do not will our maxim to become a universal law – since this is impossible for us – but rather that its opposite remain a law universally: we only take the liberty of making an *exception* to it for ourselves (or even for just this once) to the advantage of our inclination. [G, Ak. 424; italics in text]

Here Kant describes the condition in which we believe deeply in some moral principle, believe also that everyone should abide by it, and knowingly make an exception to it in our own behavior. In this case, we evade a moral obligation, as well as the charge of personal bias in application of it, by excluding our own behavior from its scope of application.

As an example, take the principle of keeping one's promises. Again we probably all can agree that the world would be a better place if everyone kept their promises, and condemn those who fail to keep theirs. And again we may hold these convictions with special fervor about those who break their promises to us. The experience of having relied on another's word in formulating an action plan, and then trying futilely to execute it as we watch its foundations buckle is equally painful, and not only because the promise-breaker disrespects and sabotages our rational autonomy. The promise-breaker's betrayal is a more elemental withdrawal of epistemological preconditions – in whose existence we were rationally justified in believing – for the prima facie success of that action plan. Again the experience of effectively thrashing around on a rug that has just been pulled out from under one's feet is extremely unpleasant.

Yet here, too, we often violate the symmetry requirement. We let ourselves off the hook about keeping promises made to others, when fulfilling such an obligation would be inconvenient, or require a greater investment of time or resources than we want to make, or when a more attractive or self-enhancing commitment beckons. Under these circumstances we may be the ones to yank the rug – perhaps on the grounds that the original promise was not that important, or that no serious harm was done by breaking it, or that it wasn't a real promise because we secretly kept our fingers crossed. In this case, as in rationalization, we invoke a self-aggrandizing justification for why the moral principle that prima facie seems to apply in point of fact does not, which dissociates our morally anomalous action from the realm of morally significant behavior.

In this second kind of case, the self-aggrandizing principle is that the broken promise is of no consequence; it's not important enough to be a real violation of promise-keeping. The mechanism of dissociation functions by identifying something in terms of the negation of the concepts that substantively articulate our theory – in this case, our metaethical theory about the

scope of application of our normative moral principles. As Kant describes the situation, we evade the application of principle to our own promise-breaking behavior, by stripping that behavior of its status as a violation of principle – by highlighting its just-this-once spatiotemporal discreteness and its not-this-principle exceptionality. Essentially we assure ourselves that our behavior does not violate the moral principle because it is not subsumable by that principle in the first place; because it is too concrete and particularized – too unusually one-of-a-kind – to instantiate it. In this case, too, the behavior is theoretically anomalous relative to our self-aggrandizing self-conception, regardless of how often we engage in it; and we pseudorationalize it by dissociating it from the scope of moral judgment.

4.3. Aristotle, Kant and Nietzsche on Denial

Although Kant's account of denial, the pseudorational mechanism that deviates the furthest from full rational intelligibility, is the most extensive I have found, he is by no means the only philosopher to weigh in on it. Aristotle in the *Nicomachean Ethics* tells us that

> the cause of involuntary action is not [this] ignorance in the decision, which causes vice; it is not [in other words] ignorance of the universal, since that is a cause for blame. Rather, the cause is ignorance of the particulars which the action consists in and is concerned with; for these allow both pity and pardon, since an agent acts involuntarily if he is ignorant of one of these particulars. [1110b31][9]

Now Aristotle probably has in mind the following kind of case. Believing I am mixing cornstarch into the gravy to serve my dinner guests, I unknowingly lace the gravy with rat poison and kill them all. Or, alternately, the kind of case in which, meaning to dust the furniture with a hand mop, I inadvertently pick up the cat, spray it with Lemon Pledge, and proceed vigorously to wipe various surfaces with it, thereby smothering the cat in dust and Lemon Pledge by accident. In both cases my ignorance of important particulars about my circumstances leads me to perform harmful actions involuntarily. Aristotle's account does not distinguish between cases in which this ignorance is justified and those in which it is disingenuous. (Who left the rat poison improperly labeled? And surely I should have noticed that the hand mop was warm and purring?)

But I focus here on a slightly different kind of "ignorance of the particular," on which pity and pardon also may depend, but for which the inference to involuntariness is not so obvious, namely ignorance of *oneself* as a particular. This is the essence of denial in moral self-deception, in which I violate the symmetry requirement by making quite severe judgments about others' violation of some moral rule, without recognizing myself to be violating that rule in my own behavior. Because I fail to apply the concept, "violation of moral rule R" to my own theoretically anomalous violation of R, I fail to recognize my violation of R, and so am genuinely ignorant of that violation.

In Volume I, Chapter II.2.3 I suggested the partial explanation for this phenomenon that a reformed Humean conception of the self would provide. Review some of the examples of this

kind of ignorance of the particular described there: Mildred, the Machiavellian social climber, complains bitterly about the Machiavellian social climbers she must contend with – 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 his own honesty to prove his points. Mavis roundly condemns Trevor – for being judgmental. In all such cases, the agent sincerely holds a moral principle and fails to recognize his own violations of it – indeed, sometimes violating the principle in the act of denouncing violations of it in others.

An observer of the scenario may wonder how anyone can be so blind to her own faults, even while discussing them in the abstract. The oversight often seems so glaring that we may find it difficult to believe that no simple hypocrisy or self-deception is involved. But hypocrisy and self-deception both presuppose knowledge – at some level of awareness – of the truth behind the deception. However, my analysis of pseudorationality implies that in ignorance of oneself as a particular, a necessary precondition of knowledge is lacking. Failing to subsume one's own behavior as a concrete particular under the available and appropriate concepts is an example of denial of a special kind, in which it is one's own, significant intentional behavior that is lost. Rational disintegrity under these circumstances consists in an interior disjunction between what one intentionally conceives oneself to do, and the significant intentional behavior one actually performs. Without its organizing rule or principle, this intentional behavior remains – to quote Kant – "nothing but a blind play of representations, less even than a dream." [1C, A 112]

How is this type of disintegrity to be explained? How can one have in one's conceptual arsenal the appropriate concepts, yet fail to apply them to the ever-present and maximally intrusive particular at hand, namely oneself? In Volume I, Chapter II.2.3, I argued that for a Humean self, rule-blindness arose from the self's orientation towards its future desire-satisfaction, relative to which available moral principles faded or sharpened in salience according to their contingent usefulness in promoting this. But this much does not explain why available moral principles might fade out of salience completely for a Humean self; or why, for any self, desire or other preoccupations ever obscure their application to it.

Ignorance of oneself as a particular is a kind of denial that originates in a failure to grasp, at a deep level, what a universal concept or principle really is. Some people self-deceptively conceive themselves as holding certain moral principles to be universally applicable. But in fact those principles only apply within the universe fashioned from their experience and structured by their conceptual scheme – relative to which they as agents figure only inferentially, as the subject-observer of that universe, rather than as a player within it. In an important sense, they do not experience themselves as contained within a universe of many, equally real and significant particulars. Rather, the universe they experience is contained within them. For such a self-deceiver, even lip-service to the concepts of impartiality and reciprocity is a stretch, for the first-/third-person asymmetry is so radical that there really is no first person for the universal principles to apply to. All there are, are other people, operating within the constraints of that subject's universe and colored by that subject's emotional responses to them. This self-deceiver's

self is absent not as object of her ministrations, as it is for the cult member considered in Section 1.1 above; but rather as individuated subject, one among others, to whom her moral principles apply. This is what Thomas Nagel calls a *solipsistic* subject: one for whom the world of third-personal agents one subjectively experiences is the only world there is. This is an agent for whom there is no further world in which his own, subjective experiences of other agents occur; no further world in which one is no more or less a subject, no more or less a player, than anyone else. A solipsistic subject fails to conceive any of his own behavior as necessarily instantiating his conceptual scheme. His behavior may conform to this theory. But then again, it may not.

Thus denial of one's own violation of moral principles one holds to be universal indicates a solipsistic universe whose creator is structurally exempt from the concepts and principles that in fact apply only to its creatures. As Nietzsche reminds us,

> The lordly right of giving names extends so far that one should allow oneself to conceive the origin of language itself as an expression of power on the part of the rulers: they say 'this *is* this and this,' they seal every thing and event with a sound and, as it were, take possession of it.[10] …When the noble mode of valuation blunders and sins against reality, it does so in respect to the sphere with which it is not sufficiently familiar, against a real knowledge of which it has indeed inflexibly guarded itself.[11]

The specific unfamiliar sphere Nietzsche has in mind, against knowledge of which the powerful inflexibly guard themselves, is the sphere of the disadvantaged. But Nietzsche's account of denial as the mechanism by which we "inflexibly guard ourselves" against "real knowledge" has broader application. And if he is right in his analysis of naming as an expression of power (and I think he is), then the withholding of a name – in Kantian terms, biased nonrecognition, can be equally an expression of power – indeed, of omnipotence. Denial can express the power to ignore a thing at no peril to oneself. As we saw in Chapter V.6.1, powerful individuals sometimes exhibit this reciprocal connection between ignorance of the disadvantaged and ignorance of the self. Solipsistic denial is a luxury only the powerful can afford.

Politicians and political organizers are thus particularly susceptible to this type of denial, of ignorance of oneself as a particular. Here are just some of the factors that incline committed politicos to orient the conscious scope of their moral judgments outward toward the objects of their attention, and away from their own behavior. There is their deep-seated devotion to their constituents. There is their genuinely altruistic concern for those whose interests they try to advance (here I take issue with Nagel's thesis that solipsism and altruism are incompatible). There is their intellectual preoccupation with defining, refining, promoting and defending their ideologies. And there are their constant battles against unrelenting political opponents who compete with them for scarce resources. The more successful politicos are in achieving their goals, the more the quality of their experience of others encourages a solipsistic sense of being the mover, shaker, and indeed the creator of their world, rather than one of its creatures.

Thus maximally successful politicos, regardless of ideology, are often famous for an almost stereotypical tangle of conflicting character traits that are paradigmatic of this type of

rational disintegrity. They often combine an ascetic dedication to their cause with alarming and often debauched personal appetites. They model compassion and charisma toward their constituents on the one hand, and imperturbable narcissism and egocentrism in satisfying their own needs on the other. They offer inspired moral leadership, and fulsome personal corruption that often emerges after the fact.[12]

But again, politicos are hardly alone in this cognitive vice. Ignorance of ourselves as particulars exempts us from the often quite harsh moral judgments we make about others; and relieves us of the pain and despair of having to make those harsh judgments about ourselves. It thus gives us the heady satisfaction of conferring on ourselves, by implication, the honorific status of moral innocence. So this type of denial should be of special concern to only those philosophers who take seriously the project of self-knowledge as essential and important to the activity of philosophy. Not all philosophers would assign self-knowledge such a high priority. But on this analysis, a low regard for the project of self-knowledge would be part of the type of rational disintegrity I mean to describe. Thus a concern for self-knowledge is a double-edged sword that may promote self-deception in some cases and inhibit it in others. The trick is to value self-knowledge without valorizing ourselves for seeking it.

Notice that in first-person moral anomaly, neither rationalization nor dissociation nor denial involves a simple conflict between theory and practice or thought and action. Quite the contrary: they each function to *unify* thought with action, by elaborately tweaking and tinkering with thought. All in some way react to one's own violation of theoretically invested moral principle, and all attempt to redress this violation through rational means. The first case, of rationalization, biases the properties of the particular so as to invite its subsumption under one principle and discourage its subsumption under another. The second case, of dissociation, biases the scope of the principle itself so as to exclude from its application a particular that it does in fact subsume. The third case, denial, withholds application of principle or concept to a salient particular in such a way that the particular itself remains unrecognized and therefore inaccessible to conscious awareness. So all three involve an inherently *cognitive* conflict that responds to the violation in kind.

All three also involve an asymmetry between the applications of the universal principle to our own circumstances on the one hand, and to others' circumstances on the other; thus is strict impartiality violated. In all three cases we apply the principle strictly to ourselves when we are the beneficiary. But we bias it toward the self-exculpatory when we are the benefactor. In all three cases we call on reason to legitimate those mis- (or non-) applications of principle or concept, by reference to other principles that inherently conflict with the purportedly universal ones. Thus in all three cases there is an internal bifurcation between the universal moral principles we conceive ourselves to hold deeply, and the self-aggrandizing principles we apply when we violate them. Pseudorationality thus sanitizes our conscience so that we may, without self-reproach, shrink from acts of courage or generosity, and embrace acts of cowardice, malice or greed.

5. The Self as Unrecognized Particular

It is tempting to conclude that the real culprit here is theoretical reason itself, and its tendency to overreach its natural limits in its hegemonic drive for conceptual control of the self, to the disadvantage of the emotions and instincts. A different version of this criticism, familiar from Western appropriations of various forms of Hinduism and Buddhism, would be that the chatterings of pseudorationality are nothing more than the mind's fulfillment of its necessary and limited function in the self; and that the real mistake is to identify the self in toto – and thus the unity of the self – with any attempt to unify the mind. Both versions of this criticism imply that in practice, integrity and coherence can be achieved only by reducing the domination of reason in the structure of the self. Despite my sympathy, particularly with the latter version of this criticism, I reject it.

Pseudorationality promotes self-deception about one's interior rational coherence. The price is a conceptually distorted, marginalized, or unrecognized particular. So far I have been focusing on that illusion itself – on how the theoretically anomalous particulars of an agent's self-conception are distorted, diminished or eradicated when the self's internal unity is riven by the morally self-deceptive disjunctions I have catalogued. I have tried to describe the conceptual violence we do to some of those anomalous particulars in the service of this illusion; and more specifically the conceptual violence we do to ourselves, when we distort, diminish or eradicate the self from conceptual self-awareness. This is the standpoint of reason, battered by the agent's own delinquent behavior, then mended inadequately by the malpractical operations of pseudorationality. But precisely because the mind that undergoes these cosmetic surgeries is *not* identical to the self, the kind of denial involving ignorance of oneself as a particular affords us the opportunity to sneak under the radar of reason, as it were, to the standpoint of that unrecognized particular itself. From this perspective, the silence of reason in the self is no more a resolution of interior disintegrity than was the invisibility of the self to reason.

Harry Frankfurt describes as a *wanton* an agent whose first-order desires are neither evaluated nor governed by higher-order ones.[13] The basic idea can be generalized beyond the constraints of the Humean, desire-based conception of the self that Frankfurt takes for granted. An agent may be motivated by sentiment, conviction, emotion, principle, belief, or need, in addition to desire; and may act in a similarly unselfconscious manner with respect to any of these normative motivational guides. When reason is silent, she fails to subsume that behavior under any concepts, fails to identify it conceptually at all. However, this does not imply that she is unconscious in quite the sense Nietzsche appears to celebrate when he says,

> [W]ith noble men, cleverness … is far less essential than the perfect functioning of the regulating *unconscious* instincts or even than a certain imprudence, perhaps a bold recklessness whether in the face of danger or of the enemy, or that enthusiastic impulsiveness in anger, love, reverence, gratitude, and revenge by which noble souls have at all times recognized one another.[14]

Nietzsche is quite right to notice that, as was true of solipsistic denial, true wantonness is similarly a luxury of the privileged. But *contra* Frankfurt, a wanton is not necessarily at the mercy of his instincts and impulses alone. He may be fully alert and sensitive to his surroundings; and may formulate intentional objects, both of consciousness and of will. However, because he lacks conceptual guidelines for self-evaluation, he lacks the tools with which fully to differentiate himself as a subject from the intentional objects he formulates. Unlike the solipsist, the wanton does not implicitly exempt himself from the principles that govern his world. Rather, he is fully identical with their practical workings. A wanton is an agent in whom self-reflective reason and so pseudorationality are silent.

The concept of the wanton would seem to imply a negative moral evaluation of an agent whose behavior, because rationally unsupervised, is morally irresponsible.[15] For example, a wanton may be motivated by the thought that it's a jungle out there and every man for himself, to fabricate documents, enter into sordid business relationships, blackmail the powerful, etc. (think Eve Harrington in Manckiewicz's *All About Eve*) – provided only that she not name her behavior to herself. Similarly, in the many accounts of childhood sexual abuse that have begun to saturate the media in recent years, I am struck by the number of abusers who used in common the tactic of conceptual silence – of maintaining it themselves, and enjoining their victims to maintain it, during as well as after the episodes of abuse. It is almost as though not thinking about or naming what they were doing while they were doing it enabled the abusers to act, by postponing the infliction on themselves of the price of self-conscious awareness, namely guilt. And it is almost as though not thinking about or naming what was being done to them as it was being done enabled the victims to survive the abuse, by postponing the infliction on themselves of the corresponding price of self-conscious awareness, namely shame. It is almost as though a requirement of complicity was somehow to prevent the left brain from knowing what the right brain was doing.

These are the kinds of cases that harden my resistance to pedestrian Anti-Rationalist arguments against "thinking too much," or "analyzing everything;" and in favor of spontaneity, instinct, and emotion. Like Nietzsche, I suspect that beneath this disdain for our sorry pseudorational fumblings lurks the arrogance – and the unscrupulousness – of power (although unlike Nietzsche, I do not approve of this). Still, using reason as a kind of cattle prod to moral rectitude is a distinctly inferior alternative.

However, there is no necessary connection between conceptual silence and morally irresponsible behavior. So not all unselfconscious agents are wantons in need of a cattle prod. An agent may act generously, compassionately, shrewdly and well – indeed, better in the presence of conceptual silence, provided only that he fail to articulate his attitudes and behavior to himself. In some agents, conceptual silence may be a precondition for those genuinely anomalous acts of conscience, skill or bravery just as it may be a precondition for unconscionable, irresponsible or wanton behavior in others. Indeed, many artists insist on conceptual silence as a necessary precondition of creativity. They find that work can only emerge when the grip of the will is loosened and the chatter of rational analysis is silenced. But some artists, and many Eastern

philosophers then develop their insight about the benefits of conceptual silence into a finely elaborated philosophical thesis. They then defend this thesis vehemently and at length, in conversation and in writing. This is ignorance of oneself as a particular writ large.

Perhaps a less obvious form of rational incoherence involved in conceptual silence consists in the failure of all of the parts of the self to work in concert; in the necessity of pulling the plug on one of them in order to maximize the performance of the others. There seems something amiss when maximal functioning of a moral, political or creative sort requires reason to be silent; when a significant part of the self must be bound and gagged in order for bravery, compassion or creativity to flower. A similar point has been made very often about cases in which it seems necessary to pull the plug on the emotions, or on desire, in order to maximize the functioning of reason and the intellect. All this may be true of us in fact. But I question whether it must or should be true, even in the non-ideal case.

6. More on Moral Integrity

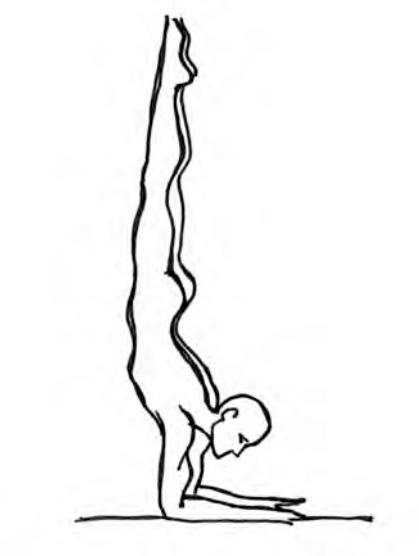


Figure 8. Pincha Mayurasana

Springing up into *pincha mayurasana*, I suddenly realize I am balancing in *pincha mayurasana*, and topple to the ground. My recognition of what I am doing undermines my ability to do it, as though my intellect were an unwelcome intruder in the intimate theater of my personal agency. Similarly, your innocent observation to your spouse that you are the family's primary economic support may bring that arrangement, and indeed your marriage, to an end – as though your intellect were an unwelcome intruder in the only slightly less intimate theater of your marriage.[16] On the other hand, it is precisely the thought that pocketing the unclaimed wallet is stealing that motivates us to return it unransacked to the police. Again recognition of what one is about to do undermines the ability to do it. But in this latter scenario, reason and the

intellect play the role of schoolmarm or cop or cattle prod, rather than party pooper. In all of these cases, rational recognition of the self – that is, self-awareness – sabotages its theoretically anomalous inclinational expression in action. Sometimes this seems a good thing, sometimes not.

It is never the best thing. Strictly speaking, there can be no best-case scenario where political machinations are necessary in order to secure the exercise of basic rights, a just distribution of social and material resources, and fundamental self-respect. This is the environment that breeds conflict between moral principle and the needs and temptations of literal self-preservation – and so the multiple operations of pseudorationality. This is the non-ideal reality in which all of us are trapped.

Still, we can imagine a different scenario – flawed in comparison to the morally integrated agent described in Chapter VI.7.3, but an improvement on the pseudorational shambles we try to piece together most of the time nevertheless. In this alternative scenario, the requirement of rational intelligibility functions not as a barricade against our unethical inclinations but rather as a fine-grained filter of them. Appropriately subtle and detailed conceptualization of our diverse emotions, impulses, and desires, deeply embedded in the structure of the self, makes psychologically harder the nonrecognition and ignorance of the self on which first-person moral anomaly feeds. Our highest-order disposition to literal self-preservation creates and reinforces a coherent and nuanced network of moral concepts and principles that strengthen and extend the capacity for self-criticism, and so discourage the unethical corruptions of power. In this scenario, rational analysis does the legwork of strengthening the interconnections and distinctions among moral concepts, and of disseminating these into general use through dialogue.

Similarly, in this alternative scenario, the requirements of horizontal and vertical consistency over time function not as an inhibitor of our intuitive talents, but rather as a welcoming structural support for them. This same fine-grained, deeply embedded network of concepts and principles identifies and alerts us to inchoate creative impulses, and enhances our receptivity to them. Practical experience then does the legwork of demanding and refining our attention to the nuances and singularities of particular subjects, objects, and states of awareness, in ways that in turn refine our grasp of them. The resulting self-conceptualization sharpens our awareness not only of form and idea, but also of the physical subtleties of *pincha mayurasana* as we are experiencing them. This augments rather than undermines our mastery of it.

Thus the discipline of thought exerts pressure on the formation of the self through the cognitive discriminations of reason, while the disciplines of practice exert it through the perceptual discriminations of the concrete particulars reason subsumes. Theory and practice mutually reinforce the extension of practice into new, challenging and unfamiliar domains of particularity, and the extension of theory into new, challenging and unfamiliar domains of abstraction.

Under these circumstances, moral integrity consists in a simple disinclination to first-person pseudorationality. It disinclines us to lie to ourselves about what we are doing, or why.

Our deeply held convictions inform our principles, our principles motivate and guide our actions, and our actions express our convictions. There is an internal coherence – in the best case, harmony – among our beliefs, our emotions, and our actions. This does not mean we never experience internal conflict – for example, between our beliefs and the impulse to self-aggrandizement. It means that when we are internally conflicted, we know we are, know what the issues are, and see the trade-offs clearly. Our self-respect does not depend on denying or dissociating or rationalizing or excusing actions we clearly recognize to be inexcusable; so we are not tempted to debase or misrepresent our core convictions to ourselves in the service of getting ahead – and thereby distort our perception of ourselves, our options, or their consequences.

Seeing clearly, free from the sophistries of first-person pseudorationality when we are tempted to violate our principles, fortifies our self-respect, and – simultaneously – a strong sense of humility; and these reinforce our interior clarity. *Self-respect* means that we have it within us to acknowledge mistakes or flaws without plunging into self-hatred or depression; that we can draw on interior resources of self-worth in order to maintain our dignity, without deluding ourselves that we are perfect. *Humility* means that we can make amends for those mistakes without feeling ashamed; that we can learn from them without losing value in our own eyes. Integrity, interior clarity, self-respect, and humility mutually reinforce one another through the sheer pleasure of heightened self-knowledge, and strengthen the self to withstand threats to its internal unity.

Moral integrity thereby nourishes intellectual and psychological freedom, for it enables our principles and convictions to emerge into our awareness from a part of ourselves that lies beyond the limitations of our self-conception; and that is uncensored by that part of our self-conception that packages our subjective self-expression for public consumption. It means that our curiosity to know and understand – ourselves, our environment, our relationships – is not stifled or constricted by guilt, shame, or fear. Intellectual and psychological freedom does not have much to do with self-assertion and even less to do with personal identity or self-indulgence. On the contrary: it is the ability to rise above the narrow constraints of the subjective ego-self, to see and investigate and understand it from a reflective distance, and to be able to use our own personal pet humans (i.e. our bodies) as instruments for being or doing whatever our principles and convictions tell us is then required – by the circumstances, by our own imperatives, or by intuition. This brand of freedom is inherently connected to the pleasure of self-transcendence, and so to the pleasure of freely acknowledging our own imperfections – with humor, compassion, severity, and accountability.

So moral integrity in tandem with freedom in thought and action is a powerful combination: It means acting in unity and inner transparency from drives and motives that lie above and beyond the blinkered perspective of the ego, according to uncorrupted principles and concepts that we deeply believe in and that inspire our action and clarify our perception, and that are unsullied by fear of public disapproval or ridicule or punishment or retaliation or failure. Moral integrity plus freedom in thought and action protects us from this kind of fear because

whenever it threatens, we see the trade-off clearly: each time we capitulate, we break our own spirit, piece by piece, one minor fracture at a time. We shatter that internal state of grace to which all other goods are subordinate as we navigate through our lives. Moral integrity, and the untrammeled freedom it nourishes, inoculates us against such self-inflicted damage. Thus we do not need to achieve the distant rational ideal of full horizontal and vertical consistency over time in order to be naturally disposed toward it. We need only the courage to choose the epistemic uncertainty of rational intelligibility to the chimera of certitude that pseudorationality represents.

7. Why I Ought Not Spend My Evenings Howling at the Moon

The possibility of preserving rational coherence in that non-ideal case in which we sacrifice certitude for interior integrity provides the basis for a detailed solution to the problem of rational final ends raised in Volume I, Chapter VIII. Chapter III of this volume laid the groundwork for such a solution, by stipulating rationality criteria a highest-ranked alternative must meet in order to count as a genuine preference. But I acknowledged there that the concept of a genuine preference does not rule out the *de re* existence of cyclical selection behavior. Therefore it does not rule out the possibility that no actual agent ever chooses rational final ends. I do not rule out that possibility here, either. But I do offer some reasons for its improbability.

The Kantian conception of the self developed in this project so far treats the self as a natural phenomenon, in many respects comparable to other natural phenomena we encounter. Like the latter, it is causally determined and shaped by forces – psychological, social, environmental – over which no one individual has any significant degree of control. As we do to other natural phenomena, we respond to the phenomenon of the self by trying to make it rationally intelligible to ourselves in socially conditioned concepts. Like the failure of other natural phenomena, the failure of the self to conform to the concepts and principles by which we explain it provokes in us compensatory defense mechanisms of a pseudorational nature, aimed at preserving the illusion of its rational intelligibility against the reality of its inscrutable conceptual anomaly. The inevitable failure of these mechanisms can lead us to revise our thinking about the self, just as it does our thinking about the behavior of other natural phenomena, and to formulate alternative concepts and principles to which the actual behavior of the self more closely corresponds.

But here the similarity with other natural phenomena ends. For unlike them, an essential feature – perhaps the most essential feature of the self is its very disposition to render its experiences rationally intelligible. By contrast to our characterizations of the behavior of third-personal phenomena that are conceptually anomalous, we are not let off the cognitive hook by dismissing our own theoretically anomalous behavior merely as, say, random rather than causal, or biologically deviant rather than stereotypical, or statistically improbable rather than likely. Instead, the inevitable failure of our pseudorational defense mechanisms to sustain the illusion of rational intelligibility disposes us, in the case of the self, to recognize our behavior, specifically, as irrational, i.e. as incoherent and therefore a harbinger of ego-disintegration; and so to reform our

behavior accordingly. The disposition to be rational may, in the end, win out over the dogmatic desire to be right. Thus the self is unlike other natural phenomena in that its interior resources for altering its own behavior patterns are identical to its disposition to understand them. And this disposition itself, which I have described as a disposition to rational intelligibility, is in turn identical to our highest-order disposition to literal self-preservation.

This point bears repeating: In practice, we are disposed to modify and reform irrational reactions and behavior in light of the ideals described in Part One, not through conscious inspiration; but instead by the hard-wired disposition to literal self-preservation. Despite the pseudorational exertions of self-deception, the very real threat of ego disintegration often pulls us back from the abyss of rational unintelligibility.

Now this highest-order disposition to rational intelligibility – i.e. to theoretical rationality – imposes an upper limit upon the proliferation of lower-order concepts and principles constitutive of the Kantian conception of the self, and so solves the problem of self-evaluation posed in Volume I, Chapter VIII.2.1. For the ascent to *n+1*-order concepts and principles from which to evaluate the *n*-order dispositions and behavior of the self are finally subject to the requirement that all such *n+1*-order concepts and principles succeed in rendering those dispositions and behavior rationally intelligible in the sense explained. But to demonstrate their rational intelligibility is to provide an authoritative justification for maintaining them. For it answers the question of why we ought to behave in a certain way by demonstrating that it is in accord with the requirements of theoretical rationality to do so. To then ask for reasons why we ought to do what it is demonstrably rational to do presupposes that in fact we ought to.

Thus contra Frankfurt, Williams and Rawls, there is in fact good reason why I ought not spend my evenings howling at the moon, whether I desire to or not, and whether that desire is a centrally definitive ground project or not. That good reason for not making a regular habit of howling at the moon is also a good reason why I ought not cultivate the intrinsic desire, at the highest order, to do so. This is that I have an idealized, coherent self-conception that includes a concept of what it means to be and to behave like a human being, with which howling at the moon is inconsistent. This concept is motivationally effective for me in that it disposes me to pick out, correctly identify, and evaluate instances of characteristically human behavior as such, to form justified expectations about my own and other people's behavior in light of it, and unreflectively to conform my own behavior to it. As Kant observes,

> [N]o single creature in the conditions of its individual existence coincides with the idea of what is most perfect in its kind; just as little as does any human being with the idea of humanity, which he yet carries in his soul as the archetype of his actions …[1C, A 318]

This idea of humanity forms a part of my self-conception that is more inclusive and cosmopolitan than my socially instilled, honorific self-conception of appropriate emotional reaction, and even more so than my morally inflected self-conception of acceptable interpersonal behavior. For it includes all of the transgressive but familiar and predictable human behavior that those latter two are designed to discourage. So to violate it, I must do much more than react with

inappropriate emotions or act unethically. There is no plausible way to draw a weak, value-neutral and widely acceptable criterion of rational final ends so narrowly that it excludes immoral or selfishly self-interested behavior. To violate my idea of humanity, I must behave in such a way that is genuine unrecognizable under the very weak and inclusive concept of human nature. That is, I must enact or become a genuine conceptual anomaly myself.

Of course, like most human beings, I do have the capacity to violate this idea in my own behavior – by spending my evenings howling at the moon, or counting blades of grass, or trapping and eating flies, or repeating the word "and" continuously from dawn to dusk, or dunking my clothes in a vat of warm lemon pudding before donning them for work each day. But if I am socialized into any human community in the usual ways, I lack the disposition to do any of these things. To then spend my evenings howling at the moon despite this would be to violate my own rationally intelligible self-conception, i.e. my conception of the kind of creature I am. It would force me to deny, rationalize or dissociate myself from my own behavior, in order to preserve my self-conception as a human being.

But these pseudorational self-defenses would ultimately fail. I could not for long deny or ignore the fact that I made a regular habit of howling at the moon, without provoking all the attendant difficulties that amnesia or multiple personality disorder tend to bring. And to what rationalization could I appeal to restore intelligibility to my conception of what I was doing? – That everyone has their harmless idiosyncrasies, perhaps? This appeal would fail to convince because as a matter of empirical fact, the range of behavior we recognize under the rubric of "human idiosyncrasy" does not extend this far. Of course our conception of human nature responds flexibly to the variety of circumstances and ways in which human nature develops. Nevertheless, it is sufficiently circumscribed that we recognize a genuine conceptual anomaly when we see it. That is, we differentiate such behavior from our conception of characteristic human behavior. But where the anomaly is first-personal, I as the anomalous agent then would be self-defensively compelled to dissociate my own identity as a human being from the actual actions I performed. Then I would be compelled to choose: between retaining my humanity by disavowing my own agency, and retaining my agency by disavowing my humanity. That I would in either case effect such a radical incoherence within the self is why it would be irrational for me to spend my evenings howling at the moon. Some such first-person conceptual anomaly is so radical that the demands of literal self-preservation exclude it even for motivationally ineffective intellects.

Now the perspective of rational intelligibility from which we are disposed to survey, evaluate and organize the lower-order cognitive, affective and conative components of the self may not be the perspective of our explicit self-conception. For if we are without illusions about the degree of rationality we are in fact able to attain, we may disavow any conscious commitment to rationality, as does the Anti-Rationalist. This may lead us, as it does the Anti-Rationalist, to reject the rational perspective as impersonal and detached from everything that gives our lives meaning. But I am inclined to dismiss this stance, too, as an instance of pseudorationality that is

ultimately incoherent. For without an overriding disposition to rational intelligibility, however involuntary, our lives could have literally no meaning, and in practice we are compelled to recognize this. A failure of rational intelligibility is a failure of comprehension, a lacuna in our accounts of ourselves, others, and the world at large. A failure of comprehension in turn signals our irradicable alienation from the object under scrutiny, i.e. the admission of the opaque and inexplicably anomalous into our conception of reality. This conflicts with our most basic instinct of literal self-preservation. For typically constituted human beings, the disintegration of the self is psychologically equivalent to the death of the self, and this is a state against which we protect ourselves at all costs. To be at once the agent of disintegration and also the self that tries to evade it is psychological anathema. The Kantian conception of the self I am spelling out in this project acknowledges and accords pride of place to this fundamental fact about us.

Endnotes to Chapter VIII

[1]See the footnote to 1C, A 551; and the further elaborated claim at G, Ak. 407-408. Also see Kant's description of a brand of self-deception at G, Ak. 424-5, and compare it with his characterization of man's natural propensity to evil in R, Ak. 32-34. For further remarks on the inevitability of self-deception and the inscrutability of our own motives, see R, Ak. 20, 38-39, 50, 62-63, 75-76, 83, 93, and 98-99. I am indebted to Henry Allison for pointing out to me the importance of Kant's preoccupation with self-deception.

[2]I doubt the difficulty of imagining alternatives to this way of thinking about oneself. For example, one might derive a great deal of self-esteem from being an academic, because one enjoys teaching and research, and believes one can make a valuable social contribution by engaging in them, without thereby supposing that academics, and so oneself, are any more important or valuable in the total scheme of things than janitors or secretaries or postal clerks.

[3]Also see Amelie O. Rorty, "Belief and Self-Deception," *Inquiry 28* (1972), 387-410. Rorty has since repudiated this view.

[4]Of course there are further, large questions about whether or not, in the absence of vicious tastes, one can be said to love a person one recognizes as unregenerately bad; and in general in what our commitment to recognizably and incorrigibly morally flawed others consists. I am indebted to Brian McLaughlin for this example.

[5]Trans. Richard Howard (London: Calder and Boyars Ltd., 1969).

[6]See David Hume, *A Treatise of Human Nature*, Ed. L. A. Selby-Bigge (Oxford: The Clarendon Press, 1968), Book III, Section III, p. 603; John Rawls, *A Theory of Justice* (Cambridge, Mass.: Harvard University Press, 1971), Chapter VII, Section 64, p. 417; and Richard Brandt, *A Theory of the Good and the Right* (New York: Oxford University Press, 1979), Chapter I.1, pp. 11-13; Chapter VI, 111-113.

[7] My gloss on dissociation owes much to John Wilson's "Freedom and Compulsion," *Mind 67* (1958), 29 – 60; and to Harry Frankfurt's "Identification and Externality," in Amelie O. Rorty, Ed. *The Identities of Persons* (Berkeley: University of California Press, 1976); although I am not in final agreement with much of what they have to say.

[8] Two well-researched anatomies of this exercise in rational unintelligibility are Philip J. Hilts, *Smokescreen: The Truth Behind the Tobacco Industry Cover-up* (New York: Addison-Wesley Publishing Company, Inc., 1996) and Richard Kluger, *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).

[9] Aristotle, *Nicomachean Ethics*, trans. Terence Irwin (Indianapolis: Hackett, 1985).

[10] Friedrich Nietzsche, *On the Genealogy of Morals*, First Essay, Section 3, 26; in *On the Genealogy of Morals* and *Ecce Homo*, Trans. Walter Kaufmann and R. J. Hollingdale (New York: Vintage, 1967)

[11] *Ibid*., 37.

[12] Here I am not thinking of the politician you probably think I'm thinking of. In fact I am alluding to Mao Zedong and the revelations about his personal habits that emerged after his death. I was particularly intrigued to learn that he never, ever brushed his teeth. I tried to imagine the effect of his rotten, green, malodorous smile on the other powerful world leaders with whom he consorted, and the desperate attempts they must have made to convey to him the diplomatic unacceptability of his total contempt for dental hygiene. My imagination failed me – just as, I concluded, their attempt to penetrate Mao's imperturbable solipsism must have failed them.

[13] Harry Frankfurt, "Freedom of the Will and the Concept of a Person," *The Journal of Philosophy LXVIII*, 1 (January 1971), 5-20.

[14] Nietzsche, *op. cit*. Note 11, 39.

[15] Thus Frankfurt says about the wanton,

> Nothing in the concept of the wanton implies that he cannot reason or that he cannot deliberate concerning how to do what he wants to do. What distinguishes the rational wanton from other rational agents is that he is not concerned with the desirability of his desires themselves. He ignores the question of what his will is to be. … he does not care which of his inclinations is the strongest. (11)

Frankfurt later issues a caveat that "a person's second-order volitions [do not] necessarily manifest a *moral* stance on his part toward his first-order desires" (13). But of course this does not imply that his indifference toward the worth of his desires is not susceptible of moral evaluation from a third-person perspective. The terms "person" and "wanton" are themselves normative, and valorize and derogate respectively; we rightly think a person ought to care which inclinations she acts on.

[16] See Ellyn Spragins' description of this interesting pathology in "When The Big Paycheck Is Hers," *The New York Times* (Sunday, January 6, 2002), Section 3, 8.

Chapter IX. "Ought"

In Chapter VIII I described in a general way how pseudorationality might function to square our morally derelict behavior with the constraints and demands of our favored moral theories, so as to eradicate any horizontal or vertical inconsistencies between them. In this chapter I sharpen that description with an analysis of the linguistic relation between that behavior and normative moral theory itself – more specifically, between that behavior and the requirements of conduct exacted by the ideal descriptive moral theory *K* described in Chapter V.5.2. There I was concerned to sketch the Hempelian structure and descriptive status of *K*. I argued that normative moral theory in general and Kant's moral theory in particular was descriptive of ideal rationality and so contained no "ought." This argument made it easy to see how normative moral theory might be continuous with explanatory theories in the social and physical sciences; and more importantly might be integrated into our informal theorizing about the world and ourselves in general. In the first *Critique's* Resolution of the Third Antinomy, Kant also observes, however, that no "oughts" are to be found in the sensible world of nature either [1C, A 547/B 575]. By arguing in Chapter VIII that in the non-ideal case, we usually act as we please, often in violation of such a theory and so in violation of our honorific self-conceptions, then pseudorationalize our moral derelictions so as to ensure the illusion of conformity to this descriptive ideal, I seconded Kant's observation in this instance as well.

If the "ought" is to be found neither in ideal descriptive moral theory nor in non-ideal descriptive reality, it would seem that moral prescriptions, commands or imperatives find application neither in the ideal nor in the non-ideal case. However, this conclusion would be premature. In this chapter I invoke the pseudorational functions described in Chapters VII and VIII to ground a linguistic analysis of "ought" as a functional intermediary between the ideal morality of Theory *K* and the reality of our imperfect attempts to conform to it. One implication of this view is that there is no distinctively normative realm, either conceptually or metaphysically. Norms can be decomposed into descriptive principles and the non-ideal, actual behavior they guide. Our magnanimous moral ideals and parsimonious moral behavior are all we have to work with.

Section 1 introduces the first of three causal factors that inculcate a personal investment in an ideal descriptive moral theory such as *K* in the process of socialization. Section 2 invokes this factor – the authority of fact – in order to explain the particular conative force and linguistic peculiarity of commands. Section 3 introduces the second two causal factors – the authority of consensus and reward – that reinforce this personal investment, and so the potency of *K* as a lens through which we perceive ourselves and our relations to others. Section 4 discusses some of the causal factors that undermine our personal investment in the truth of *K*, and call into question its explanatory adequacy. Section 5 brings this account to bear on an analysis of imperatives – i.e. sentences containing the word "ought." I distinguish imperatives from commands in terms of the degree of confidence each implicitly ascribes to the truth of *K*; and argue that "ought" has the

same meaning in the moral context as it has in explanatory and predictive contexts, in which it expresses a relation to an idealized descriptive theory whose empirical veracity is in question. I call this the "ought" of *tentative expectation*. Section 6 addresses some apparent counterexamples to this analysis, and Section 7 dissects a range of attitudes toward the truth of *K* – from pristine innocence to thoroughgoing moral corruption. Finally, in Section 8 I apply the foregoing analysis of "ought" to the case of the whistle-blower. I offer a justification for the whistle-blower's in-practice allegiance to *K* in her own behavior, even when surrounded by moral corruption; that is, even when punishment, betrayal, danger, or death is the likely alternative. In closing I enumerate some further causal factors that may weaken or strengthen our ability to meet the whistle-blower's challenge to our moral complacency.

1. The Authority of Fact

Under what conditions might we develop a personal investment in at least the lower-level empirical generalizations (A.1-4.) of *K* delineated in Chapter V.5.2? In the beginning stages of the process of socialization in morality and etiquette, our parents or guardians do not ordinarily tell us what we ought to do. Instead, we are told, and shown, what *is* done – by our parents and relatives, friends, authority figures, everyone of importance to us: that one does not eat one's peas with a knife, for example; or that we are fortunate to have enough to share with those less fortunate than us; or that one says, "Thank you," upon receiving a gift; or that adults can be relied upon to keep their promises. The process of socialization includes, *inter alia*, elementary schooling in a culturally transmitted theory of what social reality is, not what it ought to be. Nevertheless, this theory of social reality is an ideal one. For, as we later find out, not all, or even most people meet its description in their behavior.

If this theory in truth describes an ideal rather than an actual social reality, what social forces motivate us to make a personal investment in it – as descriptive of our self-conceptions, and so as prescriptive of our actions? Our instinctive childhood personal investment in adults on whom we depend leads us to theoretically invest in the concepts, beliefs and practices they instill in us. Those authority figures endow a theory such as *K* with at least three sources of epistemic and conative authority whose influence persists throughout our adult lives. First, there is *the authority of fact*. These conventions of morality and etiquette are represented to us as the reality, and the only reality. Conscientious and loving parents scrupulously screen their children from "bad influences," and work hard to do, not just say, what they want their children to do. To have reached adolescence in an environment in which others, particularly adults, have, through the example of their own behavior, successfully transmitted this idealized theory of social reality to one is practically definitive, in our culture, of having had a good upbringing, and it is one most parents strive to give their children.

Relative to this social ideal, deviations are perceived as troubling harbingers of unreality, to the extent that they are consciously perceived at all. It is a measure of the strength and endurance into adulthood of this ideal as a factually well-confirmed theory that some adults may

have difficulty in perceiving their own or others' deviations from it. For example, if it is part of one's ideal social theory that people are courteous and kind to one another, it may take one a long time before one realizes that someone has made a remark to one that was clearly intended as a slight. If one's personal investment in this theory is particularly strong, one may never consciously realize it at all. Instead, one may instinctively deny the remark, suppressing it from consciousness altogether. Or one may rationalize it, magnifying those properties of the encounter that invite interpretation of it as a particularly sardonic joke, or abstract observation without personal application; and minimizing those properties that unmistakably identify it as a slight. Or one may dissociate it, relegating it to the status of an unintelligible utterance without connection to any of those that meaningfully govern one's interactions with the offender. Each of these pseudorational strategies functions to prevent the intrusion into one's morally theory-laden view of theoretically anomalous data which would tend to disconfirm it. This is why such deviations count as theoretical anomaly relative to this theory, even though they may not be genuinely conceptually anomalous in the cosmopolitan sense. These same strategies may be elicited as well by one's own, first-personally anomalous behavior toward others, and for the same reasons: One simply may not be able psychologically to acknowledge the fact that one has made a hurtful or insulting remark to another, if one's belief that people do not behave that way is sufficiently deeply entrenched.

The authority of fact also can be illustrated by a specifically moral example. Consider the feelings of shock and disorientation consequent on being betrayed by someone you have genuinely trusted. Of course you also feel resentment, outrage, perhaps shame at the betrayal. But as is true when a promise is broken, there is an elemental sense of disbelief, a feeling that reality has shifted under your feet that these latter moral sentiments presuppose. Your personal investment in your conception of the person as trustworthy may delay the onset of these feelings of disbelief and disorientation, until the reality of the betrayal is simply inescapable. This is the type of situation that prompts the pseudorational strategies of rationalization, dissociation, and denial in order to keep one's acknowledgment of the betrayal at bay. And the effects of these strategies may be exacerbated if one's personal investment is not only in one's idealized conception of moral reality, but in one's idealized conception of the betrayer, and of oneself as a perspicacious judge of character. The same considerations apply, for the same reasons, in the event that one is the betrayer: One may rationalize one's betrayal, by minimizing one's obligation to keep trust with the betrayed; or dissociate one's betrayal, by telling oneself that one did not realize what one was doing; or flatly deny to oneself that any such betrayal took place. The authority of fact, then, disposes us to preserve our ideal descriptive moral theory as a realistic and factually well-confirmed one; and to pseudorationalize any evidence, including first-person theoretical anomaly contributed by our own behavior, that undermines it.

Not just any set of descriptive principles can receive the authority of fact. If the principles intended to describe an ideal social reality are internally inconsistent, or are seen to apply only at some times and not others, then these principles will fail to constitute an

identifiable ideal, and fail to carry authority for the child to whom they are conveyed. For example, take a child who is brought up to believe on the one hand that all human beings are equal, and on the other that some human beings – for example, blacks or women – by their very nature are made for servitude and suffering.[1] Either he must sacrifice the authority of one of these two descriptive principles, or else pseudorationalize them – perhaps by denying full humanity to blacks or women, or rationalizing their suffering as a virtue while minimizing the moral significance of the harm he thereby causes them; or dissociating as irrelevant the requirement of equal treatment, and restricting his conception of equality to equality of opportunity alone – so as to maintain the appearance of their horizontal consistency. As we have seen in Chapter VIII, pseudorationality engenders a need for further pseudorationality; and this ultimately undermines the rational unity both of the self and of external reality simultaneously. So only sets of descriptive principles that satisfy the requirements of horizontal and vertical consistency over time can preserve the authority of fact.

2. Commands

Conferring the authority of fact on an ideal descriptive theory of social reality gives commands their peculiar linguistic structure. It is commonly assumed that commands such as "Keep your promise!" are interchangeable, in most contexts, with imperatives such as "You ought to keep your promise." I argue here that this view is mistaken; conveniently, Kant agrees. Kant characterizes a *command* as a "representation of an objective principle so far as it is necessitating for a will" [G, Ak. 413] – and, in particular, necessitating for an imperfect human will. But an objective principle by itself, he says, contains no imperative [G, Ak. 414]. Therefore a command, for Kant, is an objective principle that necessitates human action without itself containing an imperative.

My analysis is basically in accord with Kant's. As we saw in Chapter V.5.2, an ideal descriptive moral theory consists in principles in the indicative mood. A command is formed from the corresponding indicative merely by dropping the second-personal subject (or, in German, switching word order). Often the two moods are syntactically indistinguishable (in both languages); and a command has greater force when it is expressed as a categorical assertion or prediction of fact. One pervasive example of the use of the simple categorical indicative to issue unconditional commands is to be found in fashion copy: "Blouses have a touch of the poet, cascading over sheer pants in fine, fluid folds," Donna Karan tells us. "This is what's Right. Now."[2] Well, there's no arguing with *that*. Consider the parity of structure among the following utterances:

(1) "This room will pass my inspection by the end of the day."
(2) "You will never embarrass me in public again."
(3) "That lout will not get another chance to ruin your party."
(4) "I will never take another drink."

The syntactical similarity of (1)-(4) suggests that commands are, in the second person, exactly what resolutions are in the first and third. They do not merely enjoin certain actions. They enforce a certain descriptive theory of reality, by flatly stating the facts as predicted by that theory. I argued in Chapter V.5 that Kant's moral theory is particularly well suited to do this because it fits as an intrinsic component into a more general, descriptive conception of reality the rational intelligibility of which enables us to preserve the unity and coherence of the self. Relative to this more general conception, the objective necessity with which Kant claims a law commands us is what I have been calling the authority of fact.

Although it is theoretically open to us to disconfirm certain parts of the theory by disobeying some of its commands, we must not underestimate the psychological and conative force a theory of reality has when valorized by the authority of fact. For example, no one would deny the powerful connection between being told that one is ugly or stupid, and feeling ugly or stupid, despite one's better judgment. Similarly, when we are told authoritatively what is the case with regard to our future action, we are disposed to comply with the facts as they are presented, no less with action than with belief:

(5) Only two pages to go on this section. You will finish reading it before taking a break.

The threat implied by an authoritative command is not necessarily the threat of punishment, as many metaethicists have claimed. The deeper threat is the threat of losing touch with reality; this was the threat that Milgram's subjects could not defy.[3] To disobey the commands derived from the moral principles constitutive of one's normative moral theory is to undermine and destablize the theory, and thus disrupt the interior unity of the self.

Consider the grammatical form of a command such as

(6) Keep your promises.

I have just been arguing that (6) is shorthand for

(7) You will keep your promises.

We hear (7) as just as much of a command as (6). But (7) reveals the grammatical structure of a command to be identical to that of a prediction of fact. (7), in turn, is grammatically analogous to

(8) He will keep his promises.

(9) They will keep their promises.

And, most importantly,

(10) Rational beings will keep their promises.

i.e. roughly (A.2) in Theory *K*. But we saw in Chapter V.5.2 that statements such as (A.2), and all similar predictions of fact, are experimental inferences from *K* – the descriptive and explanatory theory of how rational beings behave under moral circumstances. If Theory *K* describes what is the case, commands describe what will therefore occur.[4] To disobey the command is (among other things) to disconfirm the theory. On the Kantian conception of the self defended here, rationality has precisely this authority – the cognitive authority to fashion a coherent and convincing conception of factual reality and to secure one's behavioral conformity to it.

However, commands differ from ordinary predictions as assertions of will differ from abdications of will to the vagaries of fate. Whereas one who utters an ordinary prediction implicitly places a wager, one who utters a command thereby explicitly means to secure the outcome. By enforcing the factual authority of a certain conception of reality, the commander intends to ensure the compliance of the commanded in preserving this conception of reality. Sentences (1) – (4) above suggest that this analysis holds even in the limiting case, when the commander is identical to the commanded.

On the other hand, though a command may come very close to preserving all of the authority and force of fact contained in the indicative, it does not preserve all of it; and Kant's characterization of a command simply as our conception of a law as necessitating our obedience is incomplete. If that were all there were to a command, we would experience the known empirical laws of nature that determine our behavior – for example, the blinking reflex, the relation between insulin production, blood sugar level and energy – as commands, too. When we speak loosely in this way – for instance, of the "territorial imperative" to defend land, property, or human relationships against perceived invasion, or of the “biological imperative” to reproduce in the face of poverty or war, or of the “social imperative” to win status – it is usually when we feel driven to do these things against our better judgment. A command, by contrast, engages our better judgment, even though our judgment may not be the best.

3. The Authority of Consensus and Reward

A second formative source of our personal investment in Theory *K* is *the authority of consensus*. The practices and conventions represented to us by parents and authority figures as part of social reality are represented to us as what everyone does, or perhaps what nice people, or people like us do. We quickly get – and retain – the message that to deviate from the ideal is to court rejection, ostracism and punishment from those whose opinions are most important to us; and as adults we instinctively inflict these sanctions on others we perceive as deviant. For example, one of the most effective and well-known devices for putting an end to a child's temper tantrum is to isolate her in another room, alone, until she has quieted down, or to send her to bed without dinner. A jail sentence, solitary confinement, and exile to Siberia are among the institutionalized adult equivalents of this method; and as Mill observed, there are of course equally effective noninstitutionalized devices as well. Since our sense of ourselves as valued members of the human social community depends so heavily on our participation in a community with others whom we esteem and emulate, the authority of consensus provides us with particularly strong motivation to realize in both our beliefs and our actions the ideal descriptive theory which that consensus has conveyed to us. This is the weight of social pressure the whistleblower resists.

Both the authority of fact and the authority of consensus carry with them rewards for accepting the ideal descriptive theory of social reality as factual: a sense of cognitive stability and of inclusion in social community, respectively. But there are, in addition, mature rewards in a

more straightforward sense that are attendant on accepting this theory, and these, too, have authority. *The authority of reward* consists in the approval, status, goods, resources, and favorable treatment bestowed on us for "toeing the party line." For sincerely avowing as true this culturally transmitted theory, dismissing any doubts, questions, or theoretically anomalous information that might tend to disconfirm it, and fashioning our own behavior in conformity with it, we present ourselves to others as increasingly reliable, predictable, and trustworthy, and view others who behave similarly in the same light. These virtues elicit the approval and rewards of those whose own projects require them, and whose convictions are the same.

By contrast, questioning the truth of this theory too closely and doggedly, or disputing it, or ridiculing it, or drawing attention too publicly to anomalous data that embarrass it, or directly and repeatedly disconfirming it in one's behavior provokes anger, disapproval, and the devaluations of status and social standing consequent on these reactions: parental reproof for being "nosy," "fresh," "rude," or "inquisitive," perhaps; or, later in life, a reputation for contentiousness, cynicism, unreliability or disruptiveness, or iconoclasm. Even in a family or subculture ostensibly and sincerely committed to the ideal of unrestricted inquiry and research, there are usually quite inflexible constraints – constraints unnecessary for the prevention of physical harm – on what we are socially and morally permitted to ask, investigate or do. These constraints protect from criticism theoretical assumptions commonly viewed as axiomatic, as foundational and necessary for the possibility of any shared understanding at all. In fact there are very few theoretical assumptions of this sort. Most function more precisely to protect the interests of those who have benefited from them, and are defended energetically by their beneficiaries for that reason.

The authority of fact, consensus and reward thus not only helps to inculcate in us the culturally transmitted theory of social reality at very early stages of socialization. It also sustains, strengthens, and further entrenches our adult habits of thought and action throughout our social lives. A measure of the motivational force of these three sources of authority is the fear and anxiety with which we invest those forbidden topics and actions that disconfirm the ideal descriptive theory. To deviate from this ideal too radically or continually is to court punishment, social ostracism, and ultimately madness.

All of this may seem to imply that failing to keep our promises should drive us crazy, or at least ruin our lives, which it ordinarily does not. If there were no sources of epistemic and conative motivation to undermine our personal investment in the ideal descriptive moral theory that asserts this practice as a universal law, then, I submit, disconfirmation of it would have these effects with much greater frequency.[5] But there is much counterevidence that undermines the authority of fact, consensus, and reward; and thereby undermines our personal investment in the ideal descriptive moral theory they support. We regularly and inescapably witness disconfirmations of this theory, not only in our own behavior, but also, more importantly, in the behavior of those whose function it is to transmit the theory to us and reinforce our sense of its importance. By violating the laws of the theory in their own behavior, these authority figures

undermine the authority of fact, consensus, and reward, and thereby our personal investment in the truth of the theory.

4. The Loss of Innocence

Consider first the effect on our beliefs and motivation of witnessing deviations from this ideal theory on the part of a parent or other esteemed authority figure. Suppose, for example, that, having promised to come hear you play your tuba, your parents do not show up for the school recital; and that, despite your pleading and reproaches, they never make it to any of the school recitals to hear you play your tuba. In accordance with the Hempelian covering law schema discussed in Chapter V.5.2, there are a number of ways in which you may interpret this fact, each of which requires modification of the theory. First, you may question the suppressed premise that your parents are rational beings. Since young children do not ordinarily have rationality criteria independent of their parents' and teachers' behavior, this is not a psychologically realistic possibility. Second, you may conclude, after some rethinking of your higher-level conception of rational motivation, that rational beings do not always keep their promises or help the needy: Since keeping their promises and helping the needy would have been the dependable and benevolent thing to do, it seems that rational beings are not necessarily dependable or benevolent. They may be capricious or self-absorbed as well. This adjustment, which decouples morality from the more comprehensive concept of rationality of which it is at best an instance, is then a first step toward personal disinvestment in Theory *K*. For disconfirmation of its lower-level hypotheses necessitates revision of the higher-level laws that explain them.

However, we do not make such adjustments in *K* quickly or effortlessly. These require that the authority with which we have invested *K* first be undermined. That it is your parents who disconfirm *K* and authoritative others who comply with them contributes to this effect. First, the authority of reward and punishment is undermined, when you observe that your own and authoritative others' condemnation of your parents' behavior is ineffectual in altering it: They fail to keep their promises, and neither your reproaches nor cajolery, nor the reasoned intervention of your school teacher, can change them.

Second, the authority of consensus is undermined, when you observe that it is not, after all, the case that everyone keeps their promises; nor even that nice people, or people like us do so. You thereby observe, first, that individuals can fail to keep their promises without being ostracized or rejected; and second, that they can do so without your wanting to ostracize or reject them yourself. This second observation is important, for it shows you, if nothing else does, that moral dereliction with respect to the ideal moral theory does not imply the divestment of love or social identification of the derelict as a member of the group. You find that you are capable of condemning your parents for their dereliction on the one hand, and of continuing to want their closeness and affection on the other.

Finally, the authority of fact is undermined, by your observation that the reality described by the ideal moral theory is not the only reality – or, perhaps, not even the primary reality; and therefore, that deviation from it does not lead to madness. This is simultaneously the expansion of a provincial theory into a relatively inclusive and cosmopolitan one; the reduction in range of phenomena that can count as theoretically anomalous relative to it; and therefore the discovery of moral temptation, i.e. that the moral course of action does not exhaust the conceptually thinkable possibilities of action, but is instead only one among many such possibilities. You discover, that is, the distinction between the complex reality that is the case and the moral ideal that you believed or supposed to be the case. Your beliefs and expectations about your parents have been violated. They are not as you supposed them to be.

The authority of fact, consensus and reward are further undermined by one's own, inevitable first-person deviations from the moral ideal described by the theory – deviations that are now no longer so conceptually anomalous as to require the pseudorational ministrations of denial, dissociation, or rationalization. Suppose, in response to your friends' avid curiosity, you betray to each of them in turn Conrad's crush on Ruby, which he told you about in confidence. The authority of reward and punishment is undermined, when you observe that you are being rewarded for deviating from the ideal rather than conforming to it: Whereas before, your friends thought of you as rather priggish, you find that you have gained in popularity among them, as well as increasing the intimacy of your friendship with each, by relaying and gossiping at length about this juicy tidbit.

Simultaneously, the authority of consensus is undermined, when you observe that your internal feelings of guilt or defilement are not buttressed by any rejecting or ostracizing behavior toward you on the part of your community; and that your peer group not only condones but actively prefers certain entertaining derelictions over strict adherence to the moral ideal. You thereby discover that retaining membership in the group is not only not synonymous with strict adherence, but rather requires what we might call "consensual deviation," i.e. deviation from the moral ideal that involves the complicity of the very community that ostensibly advocates it. This attitude of complicity *as a value* defines and fashions the social environment against which the whistle-blower struggles.

Finally, and most complexly, the authority of fact is undermined, when you acknowledge that you have, indeed, betrayed this moral ideal; that you are yourself not one of the "nice people" who always keep their promises. However, as we saw in Chapter VIII.4, it takes most adults a long time to reach this realization, and many of us never do. Despite the evidence of our own behavior, we pseudorationally continue to suppose ourselves to be the type of individual described by the ideal moral theory. The reason for this cognitive recalcitrance is the greater interior disintegrity created by first-person than by third-person moral anomaly. It is correspondingly easier for us to distinguish between our beliefs about others and their actual behavior, than to distinguish between our beliefs about ourselves and our own actual behavior.

For another to violate our moral expectations invites at best our condemnation; at worst our rejection, punishment, or ostracism of that person. This response will be of great or little moment to the other, depending on his personal investment in our opinion of him. But a person who does not in general think of himself as a good, kind, generous, trustworthy, generally virtuous individual is susceptible, if he is socialized in the ordinary way, to the continual and severe reproaches of conscience, whether or not he heeds them. If he has internalized the ideal descriptive moral theory in the first place, then to believe sincerely of himself that he generally fails to conform to it is to believe sincerely of himself that he is bad, mean, stingy, untrustworthy, and vicious. This is to sacrifice the basis of moral self-respect. It is thereby to live with the anticipation that all the punitive sanctions of the authority of fact, consensus and reward just described will be inflicted on him – and this, of course, is psychologically to inflict them on himself.

Of course each of us are aware of our failures to live up to a shared social ideal in some respects; for instance, by not being popular, athletic, smart or attractive enough. But a person who morally dislikes herself has no internal psychological resources for withstanding the unwarranted accusations of others, for their condemnation and rejection of her then merely reinforces her condemnation and rejection of herself as an unworthy individual. She implicitly concurs with their verdict of her as guilty, alien, and debased. Thus by morally disliking herself, she allies all of her psychological and emotional resources of moral judgment – anger, contempt, outrage, resentment, shame – with those of her community and against herself.

The chasm within the self created by this radical conflict between behavior and moral self-assessment demands resolution; and there are several options. For example, cooperative condemnation by the authority of fact, consensus and reward provides a readily accessible one. By ascribing to the agent a theoretical representation of her as morally unworthy, these forms of social sanction encourage her to regard herself similarly, and move her to confirm this revised self-conception in her behavior. Blanket social condemnation thus becomes a self-fulfilling prophecy, and the agent a self-professed outcast whose destroyed or debased self-conception itself honorifically affirms her further moral delinquency. Unethical or criminal behavior becomes a badge of honor that accords additional weight to her revised self-conception; and this, in turn, provides additional legitimacy for her behavior. In tandem these two are mutually reinforcing. This is the case in which, through the original impetus of social conditioning, rational autonomy develops decisively decoupled from aspirations to moral rectitude. It is for this reason that I suggested, in Chapter VII.2, that a destroyed or debased self-conception makes an agent dangerous but not necessarily irrational. It is not irrational to believe of oneself what the authority of fact, consensus and reward prescribe, nor to act accordingly even if such action violates moral principle.

Alternative, less radical resolutions of the split between behavior and moral self-assessment are no less uninviting. The danger and diminishing marginal utility of self-obliteration – for example, through drugs, alcohol, or other addictions; and the sheer discomfort

of radical moral reform through concerted and painful behavioral reconditioning both threaten either a punishing physical and psychological ordeal, or else an otherwise destructive alienation from self of large dimensions.

Thus it is not surprising that we marshal every rational and psychological resource we have, in order to avoid recognizing this split, even where there is evidence for doing so. We may, for example, rationalize our failure to keep the promise by arguing that we did not utter the performative that would make it one; and similarly rationalize away the promisee's accusation or reproach by ascribing to him a personal axe to grind. Or we may dissociate our behavior, claiming that our promise was the result of momentary thoughtlessness that no one should have taken seriously; or we may dissociate the promisee's condemnation of our lapse, by refusing him the membership in our moral community or moral authority to pass judgment on us. Finally, we may simply deny that we made any such promise, claiming that the promisee misinterpreted our words; or we may deny his reproaches, noting merely that his social skills or choice of conversational topics leaves much to be desired.

It is only when the evidence of our own failure to conform to the moral ideal is inescapable, forced upon us, that we definitely can be said to have acknowledged our own capacity for wrongdoing; hence only then that our loss of innocence is complete. This is the point at which we recognize the ideal descriptive moral theory as just that and nothing more. The increasing disconfirming evidence provided solely by the moral dereliction of others may, to be sure, require us to draw the boundaries of the ideal moral community ever more narrowly; so narrowly, perhaps, that – tragically – only we ourselves and a few trusted friends may remain within it. However, others' motives are not directly accessible to us, and therefore resist quick inferences to moral dereliction: Because another's hurtful action may always be the result of malice or stupidity (the moral variant on Davidson's principle of charity), we may avoid any such inference – for example, if we do not yet understand the concept of evil.[6]

By contrast, to be confronted, cornered into acknowledging our own moral dereliction is to acknowledge our voluntary violation of the laws that define the ideal moral community, and thereby our voluntary defection from it. It is to have to abdicate, finally and irretrievably, the belief that we and those we know are in fact among the rational beings who always keep their promises, and therefore the belief that the ideal descriptive moral theory we have accepted as true has explanatory adequacy for the actual community of which we are members. The loss of innocence, then, is the loss of identification with and belief in the reality of the ideal community described by our descriptive moral theory – with all the feelings of pain, isolation, and unreality that self-inflicted exile from a valued community brings. This is what it means to discover that we are not as we supposed ourselves to be.

5. Imperatives

I now propose an analysis of the term "ought," as it appears in moral contexts, that is based on the foregoing discussion of the loss of innocence – i.e. on the acknowledgement to

which each of us is forced at some point in our lives, that we are not as we morally supposed ourselves and others to be. The slippage between an authoritatively established ideal descriptive moral theory and our conscious deviations from it is even greater in a categorical imperative than it was in a command. Kant defines *imperatives* as “only formulae ... for expressing the *relation* of objective laws of the will in general to the subjective imperfection of the will of this or that rational being – for example, of the human will." [G, Ak. 414; italics added] Since an objectively necessitating law is a command, an imperative for Kant is a formula for expressing the relation of a command to the imperfectly rational will it necessitates; i.e. the relation of an ideal descriptive principle to the non-ideal reality of human motivation. Rather than

(11) You clean up your room.

or

(12) You will clean up your room.

or even

(13) Clean up your room.

all of which express the factual authority of the indicative mood to some extent, the subjunctive mood of the imperative in

(14) You ought to clean up your room.

expresses something less. If a command expresses our conception of a law as requiring but not ensuring our compliance, a categorical imperative expresses, in addition, our conception of ourselves as unpredictable variables whose compliance with the law is in question.

In Kant's writings, the German word usually translated in English as "ought" is *sollen*, which means more precisely "should," “shall,” or, equivalently, "is supposed to." *Sollen* appears in sentences such as,

(15) *Die soll um sechs Uhr ankommen*.

which means,

(15a) She is supposed to arrive at six o'clock.

as well as in sentences with a specifically moral connotation, such as

(16) *Sie sollen auf Ihren Eltern achtgeben*.

which means,

(16a) You should mind your parents.

In both of these sentences, the word "should" is semantically equivalent to and interchangeable with "is supposed to." We may say indifferently, “She is supposed to arrive at six o'clock,” or

(15b) She should arrive at six o'clock.

Similarly, we may say indifferently, “You should mind your parents,” or

(16b) You are supposed to mind your parents.

The choice of utterances (15a) and (16b) heighten their factual authority, especially if the epistemic word “supposed” is dropped; and so bring them closer to the expressive status of categorical indicatives, to predictions, and therefore to commands.

But the two never meet, for the subjunctive inflection of the imperative form preserves its distinction from the indicative inflection of commands. The word "should" expresses, first of all, a belief in the agent's capacity to perform the action modified by this modal verb. Second, even more strongly, it expresses an expectation that the agent will perform this action. If she should or is (supposed) to arrive at six o'clock, then we may justifiably expect that she will arrive at six o'clock. Conjointly these first two features add up to the supposition that the agent is of a certain kind: competent, responsible, and so forth; they presuppose a background theoretical conception of the kind of being the agent is which we accept as not merely true but self-evident; i.e. as deeply embedded psychologically and epistemically. But third, the word "should" expresses something less than the certain prediction contained in "She will arrive at six o'clock." It expresses our fallibility, our uncertainty that what we justifiably expect to occur actually will occur. It expresses acknowledgment that, due to unknown interferences, our justified expectation nevertheless may be false. Fourth, "should" in these sentences connotes a faint reproach, a suggestion that, in the past, the agent has failed to do what was expected of her, or that she might so fail in this instance, which explains the uncertainty we feel. Fifth, when addressed directly to the agent whose behavior is in question, "should" functions as a reminder of what those expectations are. I shall refer to this "should" as the "should" of *tentative expectation*.

My thesis is that when used categorically, commands express predictions whereas imperatives express tentative expectations. Commands appropriate the force of factual reality because they reproduce or condense the grammar of categorical and assertoric propositions, whereas imperatives disown some of that force by modally inflecting them with "ought." *Sollen* – the "should" of tentative expectation – connects an idealized, comprehensive explanatory theory of reality, structured by the requirements of horizontal and vertical consistency over time that in turn structures our conceptions of ourselves, other people, and the world, to the empirical reality of human experience. It does this by registering the slippage between the idealized theory we suppose to be valid, and the anomalous empirical data that regularly undermine it. It is because of our personal investment in this favored comprehensive theory that the categorical *sollen* expresses our suppositions or expectations about what is or will be the case; and it is because that theory, as solid and well-grounded as it seems to be, is regularly assaulted by disconfirming or deviant data that those expectations are expressed tentatively. This proposal is perfectly general in nature, and applies to the categorical *sollen* in nonmoral as well as moral, and nonhuman as well as human contexts.

But when applied specifically to human moral contexts, the categorical *sollen,* the "should" of tentative expectation, exhausts the meaning of the word "ought" as it is used in those moral contexts. It expresses an unresolvable tension between the rationally intelligible realm of moral ideals to which we feel committed – the moral ideals expressed in principles (A.1) through (D.2) discussed in Chapter V.5.2 – and the empirical reality of moral disillusionment with which we are regularly confronted. To the extent that these principles describe our idealized self-conception as rational beings, we identify with the actions the theory articulates, accept Theory *K*

as true, and experience the commands and imperatives derivable from them as carrying the authority and force of fact. But since this self-conception is idealized, our identification with them cannot be complete. We are not unconditionally moved in every instance to perform the actions the theory articulates, nor suppose without reservation the theory to be true – nor, therefore, experience its commands and imperatives as statements purely of fact.

Thus on this analysis, when we say of someone that he should or ought to keep his promise, we mean, first, that, he *is supposed to* keep his promise. Just as a postal carrier is supposed to deliver the snail mail, according to our conception of the postal service, similarly a rational being is supposed to keep his promises, according to our ideal descriptive theory of moral behavior. Here we should not be misled by any apparent difference in "emotive flavor" between

(17) A postal carrier is supposed to deliver the mail.

and

(18) A rational being is supposed to keep his promises.

Ordinarily we have a greater personal investment in (18), and feel greater disappointment and resentment when it is violated. But it is not difficult to imagine circumstances under which we might feel as strongly about (17). In that case, I would suggest, our reaction to its violation would have the same "emotive flavor."

So we who issue judgment (18) about someone *suppose him to be* the kind of agent who keeps his promises, namely a rational being in the sense defined by Theory *K*. But we also mean to acknowledge that our supposition may be false; that he may not be, after all, as we suppose him to be. This implies that we who issue this judgment also recognize the observational fallibility of our ideal descriptive moral theory, and so a measure of uncertainty as to whether our justified expectations will be in fact confirmed by his behavior. Moreover, when we say that someone should keep his promise, we indicate our awareness that he has not or might not always thus meet our justified expectations. We thereby acknowledge the possibility that our moral theory does not adequately predict his behavior, and that these lapses (and not, say, our cynicism or lack of good faith) explain our uncertainty over his anticipated performance. Thus we express vacillation between the possibilities that the agent is not the kind of rational being described by the theory, and that the theory is inadequate to satisfy the counterfactual condition for that kind of being. In both of these ways, the moral "should" expresses epistemic ambivalence. Finally, when we address this "should" or "ought" to the agent directly, we remind him of what is expected of him; of the moral being we suppose him to be. But we recognize that of course it always remains open to him to confirm or violate this moral supposition in his actual conduct. This is to suggest that the moral "ought" expresses our relation to an ideal descriptive moral theory like *K*, under the condition that we are unsure, on a given occasion, to what extent the laws of *K* are or are not well-confirmed. In order to use the moral "ought," we must already entertain the possibility that the laws of *K* do not hold universally. Only a pristine innocent believes in the moral "is."

6. Some Counterexamples Resolved

Next I consider three objections to this analysis.

6.1. Incompatibilities

First, it would seem that the agent whose character is less than sterling may elicit from us prima facie incompatible judgments, in accordance with our conflicted and tentative beliefs about her. According to the above analysis, we may judge both that she should keep her promises (if she behaves as the theory describes), and that she should not keep her promises (if her actions undermine the theory, as we suspect they may). However, the incompatibility is superficial. For there is a semantic asymmetry between the former, theory-affirming judgment and the latter, theory-undermining one. We often make theory-undermining judgments such as these:

> (19) Veronica ought to rejoice at the philosophical howler in Avery's recent article.
>
> (20) Floyd has never been able to keep a secret; why should he on this occasion?
>
> (21) Elmer shouldn't keep his promise, unless he has recently undergone moral reconditioning.[7]
>
> (22) What is this sudden change of heart? You're not supposed to be generous when it doesn't serve your interests!

In such judgments we express, among other things, the wish at the heart of all expressions of cynicism, i.e. to be proved wrong. This wish has no parallel in the former, theory-affirming judgment, where it is replaced by a hope of being proved right and a fear of being proved wrong. Because we naturally want our moral theory to be true more than we want our cynical expectations confirmed, our theory-affirming judgment that an agent should keep her promises carries more psychological weight than our theory-undermining judgment that she should not. Indeed, that most of us do not exhibit vigilant suspicion of others by carefully choosing all our words, taping all our phone calls, and photocopying all our correspondence in the anticipation of betrayal, testifies to the psychological primacy of theory-affirming judgments. The semantic asymmetry between these two types of judgment prevents a deep logical incompatibility between them.

6.2. Incorrigibilities

Second, my analysis may seem to imply, further, that even if *we* retain some commitment to the truth of the ideal theory, those we identify as cases of what we might call *radical incorrigibility*, are released from any obligation to behave as it describes. This would be a serious flaw in my account, since I argued in Volume I, Chapter IX.4 that a moral theory must be able to generate moral judgments about those who violate all of its precepts. However, there is no real conflict. Theory *K* meets this requirement – but not by generating judgments of obligation for these circumstances. On the suggested analysis, the moral "ought" applies to particular types of

actions (promise-keeping, rendering aid, and so forth). It thus presupposes our conditional recognition of the agent as a candidate for the ideal moral community. So it cannot apply to those whose actions reveal a degree of incorrigibility that conclusively places them beyond its reach. It would be not just feeble but frivolous to express our moral judgment of Hitler by asserting that

(23) Hitler ought not to have gassed six million Jews.

– as though somehow five million would have been less objectionable. The horror of Hitler's actions enlarges the focus of moral judgment to include his motives, character, and indeed his very existence. Our attitude toward Hitler is better expressed in sheer speechlessness, perhaps; or in the judgments that he was an abomination, that what he did was unspeakable, and the like. In such cases, Theory *K* implies what we might call *judgments of negative identification*, i.e. truly rational dissociative judgments that the agent does not merely *act* in violation of some one particular moral obligation, but *is* a conceptually anomalous assault on the very conception of morality that a moral theory like *K* expresses. The challenge, of course, is to retain one's theoretical investment in *K* in the face of such repeated assaults. A second challenge that I do not address here would be to identify those cases in which judgments of negative identification are actually warranted.

By contrast with the oddly feeble sound of "ought" when applied to radically incorrigible agents such as Hitler, consider its application to an innocuously incorrigible agent of the sort who might remark,

(24) I ought to stop drinking; but you know me, I'm not going to do it.

On my account, this remark is to be understood as expressing the tentative expectation that I shall stop drinking (in the first clause), conjoined with the prediction that I won't (in the third). It is a paradigm case of epistemic ambivalence, redundantly expressed. It is not unusual to entertain certain tentative expectations of oneself based upon a flattering but insecure self-conception, and simultaneously suspect in one's heart that it is a delusion. We use the moral "ought" of tentative expectation in cases of innocuous incorrigibility because the innocuousness of the dereliction undermines our belief in its incorrigibility.

6.3. Inconsistencies

Third, what should we make of the following sentence, which seems to contain contradictory propositions?

(25) The SS men are supposed to [should] shoot the soldiers, but they ought [should] not.

Does the first clause of this judgment flatly contradict the second? If "should" were being used in the same theoretical context in each, then it would. But they are not, so it is logically possible that both clauses might be true together. Does this case show that my account does not exhaust the meaning of "ought" as it is used in moral contexts? Or perhaps that "should" and "ought" are not semantically equivalent? I think not. Both clauses deploy the "ought" of tentative expectation,

but only the second relates the agents to a specifically moral theory. That is why the second has a certain force that the first lacks. The first relates the agents to a theory of the ideal Nazi, in the truth of which we have a much weaker personal investment. The hidden references can be exposed as follows:

> (26) [According to the Theory of the Ideal Nazi], the SS men should shoot the soldiers, but [according to Theory *K*] they should not.

Similar surface inconsistencies can be generated between Theory *K* and a legal theory, or an aesthetic theory, or a theory of institutional loyalty, or a theory of economic self-interest. I have argued that what distinguishes the use of the moral "ought" in a particular case is the content of the theory it is attached to, and not any inherent peculiarities of its usage across contexts. If I am right, then such inconsistencies are to be expected in the utterances of agents who have lost their innocence.

7. Innocence, Naiveté and Corruption

Is it possible to be both a fully mature and competent adult, and a pristine innocent, without pseudorationality or ignorance? I doubt it. In Section 5 I characterized a pristine innocent as one who believes in the moral "is." Those of us who ever did begin the process of disillusionment – and so of growing appreciation for the subtlety of the moral "ought" – the first time we witness the inevitable contradiction between what our parents assert to be right and what they do themselves. Pristine innocence does not survive childhood, for those lucky enough to have experienced it there.

There is, of course, a kind of cultivated, disingenuous innocence that relies on dogged avoidance or denial of the existence of immorality or moral complexity. This is not even to see things simplistically in terms of good and evil, but rather to arrange one's life so that the experience of evil in oneself or others is denied. An example would be an Anglo-American who rationalizes her refusal to visit impoverished areas or countries on the grounds that she wishes to avoid reminding their inhabitants by her presence of how deprived they are. This is not genuine innocence, because it is based on studied, deliberate ignorance of a pseudorational sort. A concomitant of disingenuous innocence is often a lack of imagination, moral insight, and sympathy for those who undergo the torments of moral temptation. To acknowledge understanding of these torments would be to acknowledge experience of them, which is anathema to a disingenuous innocent. I would side with Kant, and against Aristotle, in suggesting that someone who lacks this kind of understanding is not capable of genuinely moral conduct.

By contrast, there is ignorance of moral corruption, not in general, but as a viable alternative for oneself under particular circumstances: You see the wallet lying open, unclaimed, and stuffed with bills near the cash machine, and it simply does not occur to you to claim it as your own. This is not pristine innocence either, but rather the effect of a deeply internalized moral theory at work in the sense explained in Chapter VIII.6. For what is lacking is not the

understanding that people steal things, but rather the interpretation of your own situation as one in which considerations of personal profit take precedence over the deliverances of moral principle.

It is, however, possible in rare cases to be both a fully mature and competent adult and a naïf, as we have seen in Chapter VII.4.1. Whereas the pristine innocent is fully invested in the factual truth of her normative moral theory, the naïf is not thus invested at all. Whereas the pristine innocent has the theoretical apparatus necessary to make both positive and negative moral judgments, the naïf does not. The naïf does not thereby avoid experience of wrongdoing, injustice or harm. Nor does the naïf necessarily lack the empathy, sympathy, and modal imagination necessary for compassion toward those who are victimized by them. But because the naïf does not view moral wrongdoing through the lens of a moral theory, nor the world in general through the lens of heavily theory-laden preconceptions, he is not vulnerable to the moral disillusionment we experience upon discovering the fallibility of our moral theories – and therefore to the anxiety and ambivalence against which we then must struggle.

For within the realization of *K*'s fallibility, there is of course room for uncertainty as to whether *K* is therefore without observational support altogether – the attitude of moral corruption; or whether, on any given occasion, it may yield a well-confirmed prediction after all – the attitude of innocence reluctantly lost. The attitude of innocence reluctantly lost is that expressed by the realization that, on the one hand, we are not as we supposed ourselves to be; and on the other, we are not supposed to be merely as we are. The moral "ought" thus expresses ambivalence in our theory-laden perception of the agent to whom we apply it. We acknowledge her moral imperfections, flaws that may disappoint our moral expectations and disconfirm our moral theory on the one hand. And we stubbornly insist on retaining those expectations and applying that theory to her on the other. The moral "ought", then, is the linguistic tool of one who, despite overwhelming evidence, is unwilling to jettison once and for all his ideal moral theory as descriptively adequate to reality.[8]

By contrast with the attitude of innocence reluctantly lost, consider the attitude of moral corruption. This stands along the same psychological continuum as that of lost innocence. The loss of innocence may be the beginning of moral corruption, if we are unable to sustain a belief even in the partial or possible reality of a moral community of the kind *K* describes. To be brought finally to acknowledge moral dereliction in ourselves is to acknowledge the existence of empirically unobservable, morally corrupt motives – power, profit, personal aggrandizement – that, we now see, anyone may have. And this realization, in turn, may call into doubt the explanatory adequacy of the highest-order theoretical constructs that govern a moral theory like *K* in the first place. In identifying our moral dereliction for what it is, we not only furnish motivational evidence that disconfirms the universality of the moral motives that *K* describes. We thereby furnish ourselves with an alternative set of theoretical constructs – the motivational concepts of self-interest, power, coercion, personal aggrandizement – that may replace them, and indeed outcompete them in explanatory power. To adopt this alternative set is to interpret

ourselves and others, no longer as adhering to or deviating from an ideal descriptive moral theory, but as instantiating a realistic amoral one in all our actions. Once this alternative theory of self-interested motivation uniformly replaces a moral theory like *K* in our interpretation of our own and others' behavior, there can no longer be any real use for the moral "ought" – except, perhaps, as a vain expression of protest against paradise lost.

For this reason, a morally corrupt individual is beyond the reach of the pangs of conscience, and therefore beyond the punishing and potentially reforming experience of moral self-dislike. Since such an individual has abandoned her allegiance to the veracity of the ideal moral theory altogether, it no longer serves her even as a tentative criterion for evaluating her own conduct. She may pay lip service to this theory, by jocularly acknowledging to others that she qualifies as a "bad person" in its terms. But even as she purports to admit this freely, she trades on the jocularity of her admission to exploit others' incredulous or skeptical reactions to it. Like Aristotle's vicious man, she derives enjoyment and personal profit rather than pain or internal conflict from her corruption.

Now Aristotle's vicious man – the prototype of moral corruption – may seem to possess a certain perverse integrity, in that by embracing both amoral principles of conduct and amoral appetites, he experiences no internal conflict between conscience and inclination. It may seem, that is, that such an amoral individual could nevertheless satisfy the criteria of horizontal and vertical consistency through time, and so qualify as rational – indeed, even as ideally rational – in the sense defined in this project. However, this is not the case. Horizontal and vertical inconsistencies abound for the amoral individual. For in order to carry out her amoral projects, the morally corrupt individual must for the most part keep her corruption to herself. She must not make the mistake Hobbes' Fool makes, by announcing her violations of social covenant to the world. Rather, she in effect must become a cleverly devious psychopath, presenting to others a smooth façade of rectitude, while covertly pursuing her amoral ends. Her morally corrupt agenda requires a policy of thoroughgoing deceit of others.

But thoroughgoing deceit is not only difficult and expensive, but also inherently confusing. In order to implement this policy successfully, the morally corrupt individual must keep track of all of the fictions she promulgates, all her utterances, their implications and practical consequences, and the evidence that supports them; and she must obliterate or discredit the evidence that undermines them. Plus she must be vigilant in her efforts to ensure the internal consistency of all of the above. Finally, she must coherently embed this internally consistent fabric of complex and detailed deceptions, and their supporting evidence, in the instrumental role of promoting the covert, amoral projects with which it is prima facie in conflict. This is a lot of work. Even Hal the evil computer broke down under the weight of such a task. In reality, not even the cleverest psychopaths manage to escape detection forever. Thus moral corruption in fact creates an enduring, inner disintegrity between the amoral principles and desires that guide action, and the immediate intentions, implications and consequences of those actions themselves.

8. Justifying the Whistle-Blower

Now, finally, to put this analysis of "ought" to work on behalf of the whistle-blower. In Section 6.2 above I suggested that radically incorrigible agents who in their actions assault the very conception of morality that a theory such as *K* expresses challenge our ability to retain our personal investment in such a theory in the face of such repeated assaults. Agents who are merely morally corrupt, when their corruption comes to light, do the same. At the same time that the public exposure of a seemingly unending succession of morally corrupt individuals reassures us that justice is being sought, it also desensitizes and demoralizes us – literally – by its frequency. With the resources of global electronic and print media and the glut of information we obtain from them, we now know we are virtually surrounded by such agents. This raises the question of why any actual agent ought to meet this challenge; why, that is, any actual agent should retain any allegiance to *K* at all. Why ought we ever refrain from doing, when in Rome, as the Romans do?

This is a particularly pressing question for Kantian moral philosophers, because unlike Humeans, Kantians cannot justify short-term moral dereliction instrumentally, with reference to its long-term beneficent consequences. Kantian moral philosophers face the same dilemma as the whistle-blower: Given the yawning chasm between an ideal descriptive moral theory such as *K* and the non-ideal reality of pervasive moral corruption with which we must all make our peace, why ought such an agent ever do the right thing, knowing that this may well lead to punishment, betrayal, danger, or death? Now I have already explored some transpersonal *reasons why* such an agent might so choose in Chapter VI.8, above – genuine preference, interiority, and motivationally effective intellect foremost among them. But this much merely *explains* the particular elements of transpersonal rationality that make the whistle-blower tick. What I have not yet done is to rationally *justify* to us the whistle-blower's transpersonal choice of moral principle over convenience, gratification, profit, or safety. That is, I have not yet made the case that anyone in this non-ideal world ought to be a whistle-blower.

A *reason for* choosing as the whistle-blower does is because, as the whistle-blower would put it, it is the right thing to do. But what does "being the right thing to do" mean? The right thing to do strengthens our cognitive allegiance to a moral theory such as *K*, by providing us with as many confirmatory instances of it as possible. When we run out of inspiring biographies to read, or arrive reluctantly at Kant's sad conclusion that

> one must listen to a long, melancholy litany of complaints against humanity: of secret deceit even in the closest friendship, so that a restraint on trust in the mutual disclosures of even the best friends is counted as a general maxim of prudence in interaction (R, Ak. 33; also see G Ak. 407-8 on Kant's doubts about the existence of virtue),

our own right conduct may be, in the end, the only source of such instances that remains; for – as the free rider demonstrates – in the end we cannot rely on others, or on any actual community of others, to sustain our own interior moral conviction. By doing the right thing, repeatedly or, if possible, whenever our moral convictions are thus tested, we supply ourselves with multiple,

successive demonstrations that *K* – i.e. the right thing to do – has relevance to actual human beings, and practical application in the non-ideal case, even though we may find few such applications in the third-personal behavior we witness. The more such confirmations of *K*'s explanatory power we muster through our own efforts, the more evidence we have for *K*'s truth as a rational ideal – that is to say, that the right thing to do is not only morally but also epistemically right. This, in turn, gives us more reason to believe in our own potential as human beings to fulfill this ideal – hence more faith in human potential in general.

This is the faith that justifies our enduring expectation of right conduct in others no matter how frequently that expectation is disappointed. For after all, we then expect no more from others than what we have called forth from ourselves. By proving to ourselves through our own moral conduct that human beings are capable of such conduct, we justify a right to expect such conduct from others. Thus doing the right thing restores and strengthens our sense of identity and value as transpersonally rational agents, and our recognition of ourselves and others as capable of transpersonally rational agency. Then regardless of whether or not other people actualize this capacity, we sustain our faith in human nature by striving to exemplify through action the best of it in ourselves.

The more evidence of this sort we supply to ourselves through our own action, the more deeply in our character we instill the habit of doing the right thing. Then the easier it becomes to recognize ourselves in the principles that *K* comprises, and so the easier still it becomes to act on them. In effect, doing the right thing reinforces our faith in these principles, and our faith in them in turn reinforces our disposition to instantiate them.

Instantiating them in this way may inspire correlative or reciprocal responses from others. Or it may have the opposite effect, of calling forth yet more concerted efforts to kill the messenger: to eliminate not only that moral disposition through bribery, blackmail or social pressure; but also the principles thus instantiated, through censorship, misinformation or "spin"; as well as oneself as instantiator, through ostracism, rejection or physical threats to one's life. In any case, all of these possible consequences are ultimately irrelevant. The primary relation that anchors the justification of doing the right thing – in this case, blowing the whistle – is that interconnective, mutual support which holds between one's self, one's actions, and the transpersonally rational ideals those actions exemplify. The permeability of this relationship to external influences is a contingent function of one's success at literal self-preservation.

This self-reinforcing and self-verifying cycle of principle and conduct, thought and action, belief and behavior, recognition and responsibility is itself a moral good, i.e. the *de facto* realization of *K* that Kant describes as "willing the supersensible world." [2C, Ak. 44] We will this "world" into existence by instantiating its principles in our own actions, regardless of the corrupt behavior of others. Doing the right thing relates us to and situates us conceptually in the midst of the ideally rational community governed by *K*; and populates our actual one with its members through our own, multiple instantiations of it. By recognizing ourselves and acting as members of this ideal community, we create a safe and protected interior moral environment for ourselves

that is independent of our actual surroundings and resistant to others’ attempts at intimidation. And through the consistency of our actions, we represent this community as an invulnerable force of right conduct to those who would make such attempts. The interior value of this ideal, and of the transpersonally rational integrity we protect by so acting comes to outweigh any actual, threatened disadvantages contingent on such conduct. Unless we thus reinforce our cognitive allegiance to *K*, the landscape of exterior social reality looks so unbearably harsh, ugly and desolate that danger or death might even seem preferable by comparison.

There are many factors that may contribute to or detract from our success in retaining some such cognitive allegiance to our ideal descriptive moral theory, in the face of overwhelming evidence that disconfirms it. One is the extent to which this theory is deeply embedded in a more general explanatory theory of the world. If one's moral theory is embedded in an explanatory metaphysics that invokes the same theoretical constructs to explain other events as it does moral and immoral behavior – for example, God, or rational purpose, or cosmic consciousness, then one cannot abandon the moral part of the theory without threatening the rest of it. If this more general theory is itself deeply entrenched in one's thinking, it may be that no amount of disconfirming evidence will be sufficient to force the abandonment of the theory. One may deploy pseudorational strategies to dispose cognitively of this evidence; or one may modify the theory to accommodate it; or one may position the theory as an object of faith rather than justified true belief. By contrast, if one's moral theory is conjoined with a mechanistic and materialistic theory of the natural world the truth of which is independent of it, the latter may be more vulnerable to evidential attack, and more readily dispensable accordingly.

A second factor that affects the dispensability of the moral theory in the face of disconfirming evidence has to do with how thoroughly and how early in one's development the ideal described by the theory has been violated by parents or authoritative others. If, for example, a young child has regularly experienced abuse, witnessed family violence, or had affection and nurture withdrawn by her parents or authoritative others, a genuine personal investment in Theory *K* may never develop. Under these circumstances, she will have no evidence that *K* holds true, even within a limited realm, and so no motivation to conform her own behavior to it. Any talk of promise-keeping, helping the needy, and so forth will be quite futile, and the moral "ought" will be little more than a meaningless sound. One may have the disquieting feeling, in the presence of such a "street-wise" individual, of talking nonsense, sounding naïve, and certainly of being completely ineffectual in one's moral exhortations.

If, on the other hand, *K* is violated less severely, and later in life, then our personal investment in it will be stronger, and our allegiance to it in the face of disconfirming evidence more secure. In this case, the explanatory adequacy of *K* will remain an open question, in spite of evidence that disconfirms it. Its laws will have the status of hypotheses, and we will simply evaluate the evidence for and against it as it comes, both from our own behavior and from others'. We will often raise questions about another's trustworthiness or benevolence toward us, and be uncertain as to the veracity of our moral judgments about her. We will sometimes

deplore those impulsive – because deeply instilled – dispositions to confide in those who prove themselves untrustworthy, and feel ashamed of our suspicions and defensiveness before those who prove themselves to be friends. We will often revise our judgments about another's moral guilt, in light of increasing evidence of her moral capacity or lack thereof, and strive to understand another's motives in a way that nullifies the appropriateness of moral blame. We will strive to cope with moral temptation, and we will often be uncertain as to the outcome of our efforts. Each of these epistemic adjustments is part of the process by which we ascertain whether, or how, or to what uncertain extent our ideal descriptive moral theory applies at all. These are the conditions under which use of the moral "ought", the "should" of tentative expectation, is appropriate. But to suggest, as some philosophers have, that we might do better without our moral theory altogether, is to fail to recognize the real alternative to it – that bleak and ugly landscape of pervasive moral corruption – that already stands much too close at hand.

Endnotes to Chapter IX

[1]Cf. Lillian Smith, *Killers of the Dream* for an interesting description of the evolution of a European American child's divided consciousness under the condition of Post-Reconstruction racism in the American South.

[2] *The New York Times*, Sunday, April 4, 1993, Section I, page 5.

[3] Stanley Milgram, "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).

[4] Obviously this does not imply that command utterances must be motivated by scientific curiosity.

[5]For discussions of particular cases, see Robin Horton, "African Traditional Thought and Western Science," in Bryan Wilson, Ed., *Rationality* (New York: Harper and Row, 1970), 131-171; and Mary Douglas, *Purity and Danger* (London: Routledge and Kegan Paul, 1966), especially Chapters 2, 6, and 8.

[6]Merely to have and be able to use the concept of evil is clearly insufficient for understanding it. Children who are taught through adult behavior the "noble lie" that the ideal moral theory is the reality do not fully understand when we merely tell or warn them that it's a jungle out there; whereas if we show them through our behavior that it is, we endow them with impulses that frequently preclude development of a stable comprehension of morality *überhaupt*. Conversely, it is difficult to imagine how a hardened criminal might ever come to make this commitment without having experienced feelings of benevolence or self-respect himself.

[7] Clearly we can distinguish between theory-affirming and theory-undermining senses both of "should" and of "should not".

[8]This, I would insist, despite the fact that a full comprehension of Theory *K* as merely a theory may entail uncertainty as to whether it is ever truly instantiated in human behavior. That a set of beliefs has the status of an Idea of Reason does not preclude its epistemic ubiquity.

Chapter X. The Criterion of Inclusiveness

In this chapter I give substance to my claim in Chapter I and in Chapter V.5.2 above, that even though it is not possible to derive one particular substantive moral theory from value-neutral criteria of theoretical rationality – as both Kant and the Humeans discussed in Volume I have tried to do, only Kantian-*type* moral theories satisfy those criteria. I proposed such criteria in Chapters II, III and V, above. And in Chapter V.5.2, I extended Rawls' analogy with science by arguing that Kant's moral theory satisfies several of them – i.e. satisfies certain basic criteria for being a genuine theory: It includes testable hypotheses, nomological higher- and lower-level laws, theoretical constructs, internal principles, and bridge principles,[1] all of which satisfy the criteria of horizontal and vertical consistency over time. I argued there that Kant's moral theory is an ideal, descriptive nomological-deductive theory that explains the behavior of a fully rational being. In Chapter IX above I argued, further, that Theory *K* generates testable hypotheses about the moral behavior of actual agents whom we initially assume to conform to its theoretical constructs; that the moral "ought" is best understood as the "ought" of tentative expectation expressed in the range of uses of the German *sollen*; and that the degree to which such a theory is well-confirmed is a function of the degree to which we judge actual, individual human agents, on a case-by-case basis, to be motivated by rationality, stupidity, or moral corruption in their actions.

However, so far I have not contended that Theory *K* is the only normative moral theory, or an exemplar of the only *type* of normative moral theory, that meets these desiderata. On the contrary, I claimed that this analysis of "ought" could be made to hold for other major contenders, such as Utilitarianism or Aristotle's moral theory, as well. So there still remains unanswered the question of which of these theories is the best among the available alternatives. To answer this question, further criteria of selection that properly apply the theory to the non-ideal reality must be invoked that cull that theory or type of theory which passes this series of reality tests from those which do not; this will complete my solution to the problem of moral justification. A moral theory is justified if it meets not only the idealized criteria of rationality I have offered in Part I; but also the further, practically adequate criteria I offer in this chapter that are in fact implicit in them. Familiar theoretical criteria I have not discussed include structural elegance and explanatory simplicity; but even these do not even begin to exhaust the desiderata for a moral theory that is adequate to the complexities of the non-ideal human community.

The requirement of vertical consistency – Chapter II's (VC) – implies a criterion of *inclusiveness* that any practically adequate theory aspires to meet. The higher the order a concept or principle has in an agent's perspective, the wider the range and variety of lower-order concepts or principles and concrete particulars that instantiate it. So the broader or more inclusive the scope of the higher-order concept or principle, the larger the range and variety of actual objects, events and states of affairs that can be recognized in its terms; and the smaller the range and variety of conceptual anomaly that conflicts with it. Then the broader or more

inclusive the scope of the concept or principle, the greater the range and variety of objects, events and states of affairs it makes rationally intelligible within an agent's perspective.

For any explanatory theory, maximizing rational intelligibility and correspondingly minimizing conceptual anomaly has obvious benefits. But for a normative moral theory, inclusiveness is not merely a benefit. It is a requirement. In a scientific theory, third-person anomaly merely threatens the coherence of the theory. But we have already seen in Chapters VIII and IX that in a normative moral theory, first- or third-person anomaly not only threatens the coherence of the self and of the agent's self-conception. It also harms the moral outcast, by either reinforcing the vicious behavior the theory condemns; or else obfuscating or excluding the anomalous agent or action from moral judgment, and so from the moral system of reward and punishment by which that judgment is outwardly expressed. So although the extension of Rawls' analogy between moral and scientific theories can be carried through to a considerable degree, moral theories are unlike scientific ones in this respect: Moral theories are subject to a criterion of adequacy – inclusiveness – that itself has moral import.

This chapter proposes a criterion of inclusiveness that a normative, practically adequate moral theory must satisfy, and that redresses the outcast status of first- and third-person moral anomaly. It argues, furthermore, that among the familiar candidates, only a Kantian-type moral theory is sufficiently well equipped to satisfy it. This means that only a Kantian-type moral theory meets all the requirements of moral justification. Section 1 offers a rough formulation of the criterion of inclusiveness that applies to both non-moral and moral theories. Section 2 sharpens the formulation through application to moral theories in particular. Section 3 narrows the focus of discussion still further, to the familiar conflict over whether an act, event, or state of affairs is morally significant at all; and if so, which moral terms most appropriately interpret it. Section 4 offers a hypothetical, non-ideal example for analysis, from which can be derived general but more detailed, practically adequate criteria of inclusiveness that address the issue of moral interpretation. Section 5 provides three such criteria that recommend for inclusion in the scope of application of one's moral theory any agent who satisfies them, and argues that these are necessary supplements to the formulation of the theory that are implicit in the ideal case. Section 6 introduces a fourth criterion, equally implicit in the ideal case, that addresses not the victim of moral wrongdoing, but rather the corrupt system that denies her fair recompense; and argues for the exclusion of such a system from moral recognition. The discussion of each of the four criteria in Sections 5 and 6 eliminates those moral theories that fail one or more of them. Finally, Section 7 concludes that only a Kantian-type moral theory satisfies these four in addition to those previously discussed.

1. Theoretical Inclusiveness

In fact I do think a case can be made that moral theories of the type that *K* exemplifies satisfy standards of structural elegance and explanatory simplicity, but I do not try to make that case here. More pressing in the case of moral theory is the requirement that the theory enable us

to address *all* the available data of moral phenomena; that its scope not be restricted by ignoring, dissociating, or minimizing the existence of moral phenomena that seem to violate its higher-level laws; and therefore that it have practical application in the non-ideal case of moral anomaly. An adequate theory needs to work in practice. It can do that by restoring theoretical moral anomaly to a recognized place within the theory as an object of moral concern.

1.1. Postow's Objection

A first, rough formulation of the inclusiveness criterion, then, requires that a moral theory be receptive to all moral phenomena, i.e. that it not commit the sins of pseudorationality, detailed in Chapters VII and VIII, against events or states of affairs of moral significance that seem anomalous from the perspective of a relatively provincial moral theory. The theory should recognize as morally significant all phenomena that are in fact of moral import; i.e. all phenomena about which moral judgments appropriately can be made.

Betsy Postow objects that this requirement in turn requires "theory-independent guidance in identifying that which really is morally significant;"[2] i.e. that a requirement of theoretical inclusiveness commits us to metaphysical realism about moral entities – or, as it is called, moral realism.[3] I do not agree. To demand of a scientific theory that it be inclusive rather than provincial, i.e. that it enable us to understand all the available data of physical phenomena does not in turn require theory-independent guidance in identifying that which really is physical phenomena. To what theory-independent guidance could we possibly turn? If the argument of Chapter II is sound, what counts for us as physical phenomena – i.e. as that about which third-person physicalistic judgments can be made – is predetermined by our cognitive apparatus: our evolving rational capacity to recognize sensory data as physical objects and events that are independent of ourselves as equally physical objects. There could be no theory-independent standard that provided us with guidance in identifying physical phenomena as physical, because there is no pre-rational way of identifying physical phenomena as physical, independent of the elementary concepts and judgments by which we begin to make our experience rationally intelligible. Just how rationally intelligible we make it is a function of whether or to what extent we craft a higher-level theory on the foundations of those elementary concepts. If "physical" is itself one of those elementary concepts, then contemporary research in particle physics instructs us as to just how primitive, theory-laden, and ultimately misguided it is.

Analogously, what counts for us as moral phenomena is similarly predetermined by our cognitive apparatus: our evolving rational capacity to recognize sensory data as having moral import. Chapter IX sketched very briefly a commonsensical account of this process, and Chapter IV.8 argued that it is not inextricably tied to interpersonal relationships. There is similarly no theory-independent standard that might provide us with guidance in identifying moral phenomena as moral, because – as we have seen in considering the naïf – there is no pre-rational way of identifying moral phenomena as moral, independent of the elementary concepts and judgments – "good," "bad," "pleasant," "painful," and the like – by which we begin to make our

experience morally intelligible. And we may similarly heighten the moral intelligibility of our experience by crafting a higher-level moral theory on the foundations of these concepts. In both the scientific and the moral cases, the difference between pre-reflective but theory-laden judgments and sophisticated theories is one of degree. In both cases, once we become aware of the theory-ladenness of our experience, we can then identify the strengths and limitations of that theory, and refine or expand it accordingly. The criterion of theoretical inclusiveness recommends that we seek always and on principle to expand its reach.

1.2. Inclusiveness

Expanding the reach of a theory may require either reformulating its laws, or rethinking its application, or both. In the United States, the Equal Rights Amendment has been controversial in part because opponents see it as redundant, arguing that its provisions are already contained in or implied by the Constitution; whereas proponents see it as practically necessary, arguing that not all legitimate subjects with a valid claim to Constitutional protection are in fact afforded it. But in fact there is no necessary conflict here. It may be true both that the scope of the Constitution already implicitly includes women; and also that it may be practically necessary to make this explicit in order to actually obtain for women the Constitutional protection to which they are already entitled. The criteria of inclusiveness I propose in Sections 5 and 6 below invite a similar line of reasoning.

Consider what happens when a scientific theory fails to satisfy the requirement of theoretical inclusiveness. Thomas Kuhn does not charge its proponents with a failure of rationality. But he does argue that a crucial role in eventually subverting the authority of that theory and contributing to a paradigm shift is often played by theoretically anomalous data that the theory not merely fails to explain but also misguidedly relegates to insignificance.[4] However, the case may be made that what is involved here is, in fact, a failure of rationality – to wit, pseudorationality – of the kind described in Chapter VII. To advance a theory intended to, for example, explain the revolution of the planets that denied, dissociated, or rationalized away the importance of their axial rotation would be pseudorational because it would sabotage the explanatory power the theory attempted to claim, by rejecting available data that should influence the formulation, scope and application of its laws. Violation of theoretical inclusiveness undermines a theory's scope of practical application.

1.3. Comprehensiveness

Of course no theory can realistically claim comprehensiveness for its explanatory paradigm, even in theory, even though it may be appropriate to aspire to it. I define a theory as *comprehensive* if it is a "theory of everything", i.e. explains all the data there is or could ever be to explain. The more a theory can explain, the more comprehensive it is. By contrast, I call a theory *inclusive* if it picks out all the data relevant to its domain of explanation. The more phenomena a theory's terms, concepts, laws, and theoretical constructs can identify, the more inclusive it is and

the more candidates for explanation it offers. Whereas comprehensiveness, on this definition, is a function of explanatory success, inclusiveness is a function of explanatory scope. A theory can be fully inclusive without being fully comprehensive just in case it can identify more phenomena than it can explain. But it cannot be comprehensive without being inclusive, because its comprehensive explanations identify all the phenomena there is.[5]

Although greater comprehensiveness means more explanatory potency, complete and thoroughgoing theoretical comprehensiveness is disadvantageous because it implies the conceptual impossibility of disconfirmation, which undermines its status as a genuine theory. By contrast, it is generally better for a theory to aspire to greater inclusiveness, in order to be able to bring within the purview of consideration the new and conceptually anomalous data that are always waiting in the wings, and against which the scope and adequacy of the theory is tested. Under these conditions a conservative epistemic policy may well be the best response. However, a practice of recognizing bona fide anomalous data, the existence of which is ascertained through replication and intersubjective confirmation, as official impetus for further revision, elaboration and hypothesis-construction in the theory is not methodologically unrealistic. Certainly it is more rational than denying the existence of such data in the hope of preserving the credibility of the theory intact.

2. Moral Inclusiveness

Similarly, to advance a moral theory that, like Kant's, purports to explain the behavior of an ideally rational agent in terms of character, principles, aims, desires, etc., that nevertheless denied, dissociated or minimized the moral significance of, for example, the treatment of men and women by one another or the treatment of children by adults would be to insure the explanatory impotence and practical irrelevance of the theory in virtually every situation in which such a theory might be expected to provide guidance. This would be a paradigm case of pseudorationality. A practically adequate moral theory cannot ignore the actual data of moral experience, on pain of vitiating the formulation, scope and practical application of its laws. As an antidote to pseudorationality in the construction of a moral theory, we may therefore require of a moral theory that it be maximally sensitive to what counts as moral data; that it include all morally significant behavior within its domain of reference, and not confine its purview to simplistic injunctions to keep promises or maximize happiness. As in the case of the ERA, this may require explicit mention of subjects, or of properties of subjects, to which the original formulation of the theory already implicitly applied.

2.1. Moral Recognition

We can then require, as a criterion of practical adequacy, that the theory be sufficiently inclusive such that in the formulation of its descriptive laws and practical principles, it is practically capable of *recognizing* as morally significant all the behavior to which moral praise, blame, or acquittal is a relevant and appropriate response. For example, a moral theory that

yields applications to newly formulated specific issues, such as contemporary Utilitarianism has done with regard to the issue of animal rights,[6] satisfies the criterion of inclusiveness, but not merely by extending its reach downward to the empirical. Classical Utilitarianism, as well as the casuistical elements in Kant's moral theory both do that much. The metaethical importance of the contemporary Utilitarian discussion of animal rights is that it extends the scope of the theory outward as well, to encompass pre-existing moral phenomena, now clearly recognized as such, that were theoretically implicit in but not formerly identified as falling within the moral domain. Full appreciation for Utilitarianism's criterion of sentience would extend the scope of the theory even more widely than this – into moral territory now occupied only by the Hindu philosophy of Jainism.

2.2. Explanatory Strength

A moral theory that satisfies the criterion of inclusiveness as roughly formulated here is distinct from a theory that satisfies criteria of explanatory or practical *strength* (comprehensiveness is sufficient but not necessary for explanatory strength). A theory that has explanatory strength can generate practical solutions for new moral phenomena that the theory may not originally have foreseen. An example of a theory that satisfied this latter criterion might be a Kantian theory that, because of the conception of rational capacities built into its theoretical constructs, generated definite answers to practical moral questions that turn on weighing the importance of rationality. Examples would include the questions of whether abortion in the first trimester is justifiable (yes, because rational capacities have not yet developed), whether human stem cells or fetal tissue up to that age can be used in treating Parkinson's disease (yes, for the same reason), and whether self-monitoring and self-correcting robots of a level of cognitive complexity comparable to ours – such as Commander Data, for example – are moral agents (yes, because rational capacities are sufficiently developed). This would be an example of a theory that had explanatory strength, i.e. yielded testable hypotheses and valid inferences about agent character and action under previously unforeseen circumstances, in virtue of the empirical validity of its higher-level laws.

By contrast, satisfaction of the criterion of inclusiveness requires that the formulation of a theory's laws and principles take into account all the *existing* moral data, and not just some of them. This means that the theory's laws and principles, if insufficiently inclusive to begin with, may need repeated revision, reformulation or supplementation in order to capture an ever-increasing range of moral phenomena. They also may need to be subject to standing practical review, in order to ascertain that all agents selected by the theory as implicitly falling within its domain of moral treatment are in fact treated as moral agents. For example, a Kantian-type moral theory that specifies all rational beings as being within its domain of moral treatment may require review in order to determine whether or to what extent its laws and principles apply to fourteen-year-olds, the great apes, highly complex computers, or extra-terrestrials that, although looking like gigantic centipedes, can nevertheless play chess. Kant's own moral theory has, in

effect, been under this type of review for the last 215 years, as we ascertain that its laws and principles apply equally to the women, blacks, and Jews that Kant himself would have excluded from it.

2.3. Inclusiveness vs. Strength

A theory can have explanatory strength without being inclusive. For example, Kant's own theory might yield the valid futuristic inferences described above, yet be said to lack inclusiveness by making no provisions for the treatment of animals or the mentally impaired in its laws and principles. By confining his discussion to rational beings, Kant formulated his moral theory less inclusively than did Utilitarianism. Of course Kant's formulation of his moral theory does not rule out the possibility of supplemental laws that might describe a rational being's nonreciprocal moral obligations to children, animals, the mentally impaired, or the environment. It is not difficult to sketch a line of theoretical reasoning that implies such obligations.

Additionally, a theory can be so inclusive as to lack explanatory strength entirely, as does the psychological egoist's that all actions are motivated by self-interest, or Anne Frank's that all human beings are good at heart, from which no testable hypotheses can be generated. So explanatory strength and inclusiveness are mutually independent. A theory that has explanatory strength but lacks inclusiveness is less adequate than one that has both, because its hypotheses are vulnerable to disconfirmation by the theoretically anomalous data excluded from them. Aristotle's exclusion of women and slaves from the moral domain might exemplify this vulnerability. A theory's explanatory strength enables us to forecast the future; its inclusiveness enables us to see what is under our noses.

2.4. Disconfirmability

Note that satisfaction of the criterion of inclusiveness does not conflict with Popper's requirement of *disconfirmability*, since this is the requirement that the higher-level laws and theoretical constructs of a theory not be tautologous. A moral theory can satisfy criteria of inclusiveness and of disconfirmability simultaneously because it can be true both that the theory identifies all the relevant data and also that its explanations make inaccurate predictions. For example, a Kantian moral theory might generate practical principles that both apply to all agents who have any rational capacities whatsoever – hence satisfy inclusiveness to some degree, and also are disconfirmable by, say, an agent who fully exercises those capacities and disciplines his sensuous inclinations in the ways Kant specifies, yet regularly violates the prescriptions of moral principle. One consequence of tying his account of rationality so closely to his account of morality is that Kant rules out the possibility of a fully rational agent who is also morally vicious. This speaks in favor of the claim of Kant's moral theory to the status of a genuine theory, but against its explanatory potency.

2.5. Inclusiveness vs. Strict Impartiality

The criterion of moral inclusiveness is also distinct from the metaethical requirement of *strict impartiality* in the application of a moral theory's laws. To recapitulate briefly the discussion of Chapter VI.1, 5 and 6, strict impartiality requires that similar cases be treated similarly, without bias either towards one's own case or towards others'. But impartiality in the application of a theory's laws is compatible with a failure of inclusiveness in the formulation of those laws itself. Aristotle's moral theory, for example, may be said to apply impartially to all citizens of the polis, yet for that very reason ignores, dissociates, and rationalizes women and slaves out of moral consideration. Similarly, a theory may be inclusive in that its laws and principles identify as morally significant all behavior that in fact has moral import. Yet it may fail to treat similar cases – as picked out by the terms of the principles themselves – similarly, and may thus express bias towards a particular group, person, or set of interests in the way it is applied. This would be the criticism made by proponents of the ERA. A moral theory that satisfies both inclusiveness and impartiality both incorporates all the relevant data into the moral domain in the formulation of its laws and principles, and also accords them their due once they are there.

2.6. Inclusiveness and Moral Interpretation

The criterion of moral inclusiveness is important because only a theory that satisfies it as well as the others mentioned will be sensitive to those nuances of social interaction that are of no less moral weight for being subtle in their manifestations, and therefore no less in need of guidance by moral principle. For example, are casually disparaging jokes about a professional competitor, uttered in the presence of powerful colleagues, grounds for moral condemnation? Does an attempt to convince a partner to accept one's occasional adulteries by threatening to otherwise end the relationship and withdraw economic support count as psychological coercion? Does confiding in one's pre-adolescent offspring about one's romantic entanglements constitute child abuse? These are instances of seemingly trivial behavior that may have major moral ramifications, if they are brought within the realm of moral concern.

The question in each such instance is whether the particular act-token in question should be brought into the moral domain or not – a moral variant on Nietzsche's more general observation, discussed in Chapter VIII.5, about the power of naming. This is the dilemma, not about which of two mutually incompatible and equally obligatory acts to perform; but rather about which of two mutually incompatible and equally compelling interpretations of an act to accept: that which situates it inside, or, alternatively, outside the range of morally significant behavior. Typically, one interpretation of the act identifies it as a moral dereliction – and therefore subject to moral control, whereas the other identifies it as irrelevant to or outside the scope of moral discussion. The former interpretation presupposes a moral theory that includes this type of act within its scope, whereas the latter interpretation presupposes one that excludes it – and thus transforms into a theoretical anomaly what ought to be well within its range. Thus the dilemma is not generated by an inconsistency in the moral theory we accept, but rather is a

dilemma about *which* moral theory *to* accept in order to understand the act in question and the data of moral experience more generally. This is the issue under discussion in this chapter. By examining some of the issues involved in granting or withholding moral significance in interpreting a particular act, I try to suggest in somewhat more detailed terms what the criterion of inclusiveness comes to in the case of moral theory.

3. Moral Interpretation and Vertical Consistency

The goal of understanding the data of moral experience by subsuming it under the terms and concepts of a moral theory is distinct from that of explaining the data of moral experience. The question is not the relatively higher-level one of which hypothesis about ideally moral agent character will correctly predict the act in question as an outcome. That question can be raised only following an answer to the more basic and essential questions as to whether the act is morally significant at all; and if so, under what moral rubric it should be subsumed. Thus a resolution of the dilemma will yield us the correct, theory-laden observational term to apply to the act in question: Is it an abuse of power? A betrayal of trust? Or, alternately, is it an act of conviction? Or an affirmation of loyalty? Or is it more appropriately treated as an innocuous act, unremarkable in its moral neutrality and so inherently proscriptive of moral commentary?

That these questions are raised at all probably rules out the last-mentioned alternative. A genuinely innocuous act does not proscribe moral commentary; it renders it superfluous. The proscription of moral commentary is, more usually, a conspiratorial proscription of boat-rocking – a sure sign that moral commentary is urgently needed in order to prevent the boat from sinking and the rats from jumping ship. In order to arrive at an answer to these questions, characterizing the sequence of behaviors in morally neutral terms alone is insufficient unless there is prior intersubjective agreement on its moral significance or lack thereof – in which case the search for moral terms in which to describe it is unnecessary.

But prior intersubjective agreement does not always exist. Some people need to have explained to them what is questionable about using federal funds earmarked for low-income housing to build a luxury high-rise for personal profit. Others understand what there is to question, but conclude, in accordance with the dictates of their moral theory, that the questions can be answered without imputation of wrongdoing. We begin to discover which moral theory we actually accept in practice when we settle the question of how to describe the acts on which it passes judgment. And we may sort moral theories into those that recognize and provide appropriate sanctions for certain kinds of acts, and those that recognize and provide sanctions for different ones. We may have to begin with morally neutral terms when these other questions are at issue. But we can end with them only when all of them have been resolved.

This is not to claim that morally identifying an act is sufficient for identifying the particular moral theory that evaluates it. The data of moral experience is regularly overdetermined by the plethora of moral theories that may be invoked to explain it. For example, both Kantian and Utilitarian theories may prescribe promise-keeping, the first as an expression of

respect for rational ends in themselves and the second as a dispensable means for maximizing happiness. Similarly, both theories may agree that killing, when neither for self-defense nor for defense of one's national borders under conditions of declared war, is murder. Any choice of an observational term is, however inherently theory-laden in itself, consistent with a variety of upper-level theories that may succeed in giving it contextual coherence.[7] The term finally chosen may commit one only to an identifiable range of moral theories.

All the theories in this range may concur in condemning, or praising, or acquitting the agent for a particular act. Yet they may differ as to the practical consequences of this condemnation, praise, or acquittal. For example, three different moral theories may agree that rape is morally blameworthy. Yet one may prescribe punishment and ostracism for the perpetrator, while another prescribes punishment and ostracism for the victim, and the third prescribes no punishment to anyone because other considerations always outweigh it. We may use our responses to such examples as a guide to solving the dilemma of which range of moral theories we should choose in order to identify the correct moral interpretation of a particular act, relying on detailed refinements in the case under study, and our responses to them, in order to narrow and sharpen the particular moral theory to which we ultimately find ourselves to be committed.

In part this can be ascertained by measuring our willingness to act on the practical consequences of a particular moral interpretation the theory prescribes; this willingness is what distinguishes the whistle-blower from her co-workers. And in part it can be ascertained by gauging the explanatory power of the theory that results from excluding or including this interpretation in it. So, for example, we may discover our unwillingness to apply the relevant moral sanctions to an act we initially interpreted as morally blameworthy. In this case we can either revise our moral interpretation of the act within the theory, or jettison that type of act from the domain of the theory altogether.

Suppose the former alternative ramifies throughout the rest of the theory in such a way as to generate vertical inconsistencies. Suppose, for instance, that after discovering our unwillingness to prosecute date rape, we revise our interpretation of the act so as to excuse date rape while continuing to condemn physical assault more generally (perhaps on the grounds that the concept of a date implies a mutual presumption of intimacy). We are then confronted with a *prima facie* vertical inconsistency, between proscribing physical assault in general and permitting what would seem to be a particular instance of it, that damages the viability of the theory. In order to repair it, the dilemma of moral interpretation may be raised again: Is so-called date rape really an instance of physical assault – thus subject to moral sanction? Or is it just particularly energetic sex between consenting adults – thus (at least on some accounts) morally unremarkable? The dilemma of moral interpretation may be reiterated at increasingly higher-level laws of the theory. Thus one may also call into question whether kissing someone could ever constitute physical assault; whether physical assault itself is always a bad thing; whether bad things may not be more accurately identified as good if their consequences are; and so forth.

Alternatively, we may solve the dilemma of moral interpretation by circumscribing the scope of the theory more narrowly. For example, we may deny that date rape ever in fact occurs (perhaps on the grounds that the recipient indicates his or her desire for sex by going on the date in the first place). Or we can circumscribe the theory even more radically, by jettisoning physical assault in general as a type of act warranting moral condemnation. Thus we may fiddle endlessly and pseudorationally with the interpretative terms of the theory so as to avoid the consequence of having to prosecute date rape, finally transforming a vague but unexceptionable moral theory into a bizarre pseudorational parody of moral reasoning. In order to avoid getting stuck with a moral theory vitiated by vertical inconsistency, moral blinders, and bad conscience, we must either fashion a different theory that avoids these evils, or else rethink our unwillingness to act on our original condemnation of date rape. Only after we have solved the dilemma of moral interpretation of the particular act in some such manner does the type of moral dilemma concerned with conflicts between obligations arise.

Thus the target of scrutiny under discussion in this chapter is the moral theories we hold in reality, as revealed in our social behavior – not the abstract and idealized theories we may defend in discussion. These latter theories are inherently inadequate to the moral data because their principles qua principles cannot fully reflect the complexity of our actual moral practices. By focusing on the question of how to apply the criterion of inclusiveness in subsuming under a moral rubric acts we often assume to be morally unremarkable in practice but that are rarely addressed in metaethical discussions of moral theory, we may be able to articulate a practically viable moral theory that can be distinguished both from an impractically idealistic one on the one hand, and from the frequent deviations from any such theory that regularly prod our conscience on the other.

In settling on the morally appropriate terms in which to describe an act, we may discover not only the range of moral theories to which we actually subscribe, but also the particulars of our own personal investments in the issues under consideration. If we identify the act as a moral dereliction, condemnation or perhaps even some stronger intervention action may be called for, whereas if not, we are let off the moral hook. Being ever reluctant to assume the burden of moral responsibility, we may prefer to fiddle with the terms of our moral theory in the manner just described, in such a way as to allow us to see the act as morally innocuous, and hope that the case for that interpretation will stick. Thus, as we shall see, fixing on the correct verbal description of an act can be a case study in pseudorationality that ultimately yields its own moral strictures, for it requires us to distance ourselves from our personal investment in evading culpability – by resisting the temptation to deny clear evidence of wrongdoing, or to dissociate that evidence as irrelevant to the broader significance of the act, or to rationalize the subsumption of the act under less morally charged concepts.

Even thinking about this issue in the abstract presents this difficulty, for we may find ourselves instinctively identifying or sympathizing with one or another agent involved, and this, together with our reluctance to encourage attributions of moral responsibility to ourselves, may

influence our willingness to identify any as perpetrator or as victim. Consider, for example, the Viet Nam veteran who protested the rail transportation of chemical weapons across state lines by lying on a railroad track, and was then sued by the conductors of the train that cut off his legs, charging *him* with having caused *them* mental anguish. "Blaming the victim" is, in this as in other comparable cases – rape, wife-beating, child abuse, sexual harassment, for example, a misnomer; for to those instinctively allied with the instigator, it is obviously not the victim who is being blamed.

In this way, who counts as the victim and who as the perpetrator cannot be settled in advance of settling the question as to how the act itself is to be morally interpreted; and settling these questions in turn settles the further question of who, if anyone, is to be blamed. What is not settled thereby are the questions of just how blameworthy the perpetrator is judged to be, and what form any consequent punishment should take. Settling these further questions of comparative degree will help situate the act and the agent within a broader moral context in which other acts are weighted and evaluated in relation to this one. This process of inquiry, in turn, will help focus the boundaries and content of the particular moral theory we finally accept.

4. Test Case #3: The Great War for Control of Reality

In what follows, I begin this process of inquiry by discussing at length a hypothetical example in which the moral interpretation of an act is in dispute, in order to derive at least some of the more specific requirements on a moral theory to which satisfaction of the criterion of inclusiveness commits us. The point of the example is to explicate what I assume to be shared methodological intuitions of moral salience, and then to formulate them as more detailed elaborations of the criterion of inclusiveness offered at the outset of this chapter. One implication of proceeding in this way, which I accept, is that intuitions that directly conflict with those I formulate here as criteria of inclusiveness are based on some sort of cognitive deficit: incorrigible pseudorationality or psychopathy, perhaps. I address incorrigible pseudorationality about racism, misogyny, homophobia, elitism, and anti-Semitism in Chapter XI, following.

Because the resulting criteria are metaethical requirements on any adequate moral theory rather than substantive requirements on a particular one, they call our attention to certain recognizably moral data that must be given weight within an adequate moral theory. They do not thereby provide an answer as to how this data should be weighed within the domain of any particular moral theory, nor how particular individuals should be treated because of it. Nor do they provide substantive answers to any other pressing moral questions in which competing interests have a claim on our moral consideration (for example, to the question whether a human fetus has rights that outweigh a woman's right to control her own body). Rather, the strategy is to examine certain typical, pseudorational mechanisms by which such data are excluded, and then to derive more specific criteria of inclusiveness from them that appropriately situate these data within the moral domain. Although I conclude that only a Kantian-type moral theory

satisfies each of these criteria, this is not to deny that there might be further criteria of inclusiveness that it fails to satisfy.

The example runs as follows. *Smith* is the Philosophy Department Chairman, a full professor, and a European American male. *Vogeler* is his colleague and pal, also a full professor, and a European American male. *Washington* is an assistant professor, untenured, and an African American female.[8] Some of the remarks Vogeler makes to Washington over the course of her first semester are as follows: that Washington certainly is a hot number and must have a lot of boyfriends; that Washington only got this appointment because she is black; that Washington looks just like the sexy housemaid Vogeler's family used to have; and that Washington must learn to be more friendly to her senior colleagues if she wants to get tenure. Some of the remarks that Vogeler makes about Washington to her male graduate students and to his male colleagues are as follows: that Washington does not know the literature well enough to teach her courses; that Washington does not like men; that only effeminate wusses befriend a ball breaker such as Washington; and that Washington is going to complain to the university administration about the department's treatment of her. Vogeler in fact consults the university's legal counsel himself as to how the department can get rid of Washington without incurring a discrimination lawsuit, and brags about this to his colleagues. Washington gets wind of these events, describes all of Vogeler's behavior to Smith, and asks Smith for help in putting an end to it. Smith replies blandly that all junior faculty find it difficult to "run the gauntlet" in order to get tenure; that he has known Vogeler since college; and that Washington is overreacting, seeing offense in Vogeler's behavior where none is intended.

Clearly, Washington and Smith accept different moral interpretations of Vogeler's behavior. Washington condemns it as harassment, whereas Smith treats it as without moral import. Which of them is correct? Is Vogeler's behavior to be described as harassment, or as mere fraternal hazing? Is it possible to decide between them, or must we content ourselves with impotent musings on the subjective incompatibility of different worldviews?

That mere different worldviews are not what is at issue is signaled by Smith's calling into question Washington's competence to make a considered moral judgment. By accusing her of overreacting, of taking Vogeler's behavior too seriously, Smith does more than suggest that Washington might be mistaken, in this instance, in her moral evaluation. A mere mistake in moral judgment can be corrected with added information or further reflection on the implications and consequences of action. It is susceptible to adjustment through the application of rational procedures of information-gathering and inference. Thus it can be revised within the framework of the substantive moral theory that the mistaken moral judgment presupposes.

By contrast, if I react with vehement repugnance, upon learning of a African American man who has been beaten to death for venturing into a segregated European American neighborhood, it is because such an act violates my favored moral theory, i.e. my values. There is no mistake in judgment I have made that can be corrected by learning that this is common practice in many parts of the United States, or that the man was a drug addict, or by adjusting my

stance to reflect my probable partiality as an African American. If an unsympathetic observer suggests that I am overreacting, seeing personal malice where none is intended – perhaps the murder is intended merely as an impersonal deterrent, to keep African Americans in their place – the implication is not only that my values are misplaced; but also that my capacity for moral judgment itself is therefore impaired. By disvaluing too strongly the practice of murdering African Americans who trespass onto European American territory, the observer might reason, I am revealed to be incapacitated from passing reliable judgment on a whole host of moral issues.

Similarly, Smith's suggestion that Washington is lacking in reflective balance, or evaluates Vogeler's behavior too negatively implies that those rational processes themselves have been subverted by Washington's psychological or emotional makeup, and hence that her substantive moral theory itself is deficient. Smith also implies his own authority and competence to make such a judgment, based on his superior knowledge of Vogeler and of the tenure process, and on his greater distance from the conflict in question. Smith's response to Washington thereby raises essentially the same dilemma, about how to choose between moral theories; at the meta-level, of how to choose between choosers of moral theories: Is Washington's identification of Vogeler's behavior as harassment itself evidence that she is defective as a moral judge? Is Smith's identification of Washington as defective in moral judgment itself a testimonial to his own moral acuity? Who is to decide between Smith and Washington as to who is the more reliable moral evaluator?

This hypothetical example demonstrates that the object-level dilemma, of how to choose between competing moral theories, is not *conceptually* dependent on the meta-level dilemma, of how to choose between competing choosers of moral theories. In theory it is possible that, rather than attack Washington's credibility as a moral judge, Smith might have politely begged to differ with her interpretation and retreated from the field, thus shirking his own responsibility as moral mediator. This would have left intact the presumption of Washington's equal status as a competent player in the game of moral evaluation. But it also would have left unresolved the impasse between Smith and Washington, as to whether Vogeler's behavior was morally blameworthy or not. This impasse must be resolved if Vogeler's behavior is to be situated within the system of practical moral controls that govern the community of moral agents of which Vogeler, Smith, and Washington are all presumptive members. Otherwise the efficacy of that system itself will begin to deteriorate, to no one's ultimate advantage.

So it is not a trivial matter which interpretation of Vogeler's behavior finally prevails. Nor is it merely a matter of intellectual disagreement that Smith and Washington have different moral views of this. Their respective moral theories concur to the extent of agreeing that *if* Vogeler's behavior toward Washington constitutes harassment, Vogeler is morally blameworthy and Washington deserves vindication. Where they differ is at the crucial point of determining what overt physical behavior constitutes harassment and what does not. For example, it may turn out that Smith's moral theory groups under the rubric of "harassment" only physical abuse – pinching, hitting, rape, etc., whereas Washington's theory groups under that heading any hostile

behavior that causes her intense mental distress, i.e. emotional and verbal as well as physical abuse. Determining which of these theories is to prevail is also to determine which of these theories is more adequate to the data of moral experience – i.e. which most perceptively and inclusively identifies behavior to which a condemnatory or condoning moral response is appropriate. This is important because that theory, in turn, will determine when and where to apply the practical moral controls that return the community to equilibrium; and who has a say in deciding in what community equilibrium consists.

In this enterprise there can be only one winner, and polite talk of the subjective incompatibility of different worldviews is beside the point. If Washington is right, Smith and Vogeler are morally culpable and she is not; whereas if she is wrong, she is morally culpable and they are not. Washington's and Smith's moral theories are not just different; they are competing, and serious personal and professional consequences follow for everyone, depending on whose moral theory prevails. To fight this "war of words" is thereby to fight the Great War for Control of Reality, in which no prisoners are taken. Hence from the no-holds-barred perspective, it is perhaps not surprising that Smith attempts to undermine Washington's evaluative authority and credibility at the same time that he rejects her moral judgment. The object-level dilemma is *practically* dependent on the meta-level dilemma, because the authority and credibility of one's favored moral theory presupposes the authority and credibility of oneself as moral judge.

5. Implications of Inclusiveness

5.1. Recognition of Rationality

The practical dependence of the object-level dilemma on the meta-level dilemma itself provides a starting point for deliberation about the relative merits of Washington's and Smith's favored moral theories respectively. Although there can be only one winner of the competition among moral theories as candidates for the actual system to which the community of moral agents consistently adheres on a particular occasion, a moral theory that prevails because its proponents have obliterated, ignored, or sabotaged the credibility and authority of their rivals is no real winner at all; for it cannot command the rational assent of those rivals who continue to maintain different theoretical allegiances. In reality, Smith's attempt to devalue Washington as a competent moral judge *to her face* is a pseudorational attempt to simultaneously deny her authoritative status as a moral agent and gain her theoretical allegiance, without examining rationally the case to be made on her behalf. If he can convince Washington that her mental distress is excessive relative to the event that purportedly caused it; that that event did not in fact cause it because Washington saw offense in inoffensive behavior; and that in any case Washington's reaction is unimportant relative to preserving the collegial *status quo*, he will have convinced Washington, effectively, that she really was just "seeing things," and so that there is no moral case to be made on her behalf after all. In this instance, Smith's moral theory prevails, not

through considered evaluation of its merits, but rather through ideological reprogramming of the opposition.

However, for Smith to succeed in convincing Washington that Vogeler's behavior was innocuous hazing rather than harassment would be for him to convince her that Vogeler's behavior was appropriate, whereas her reaction was inappropriate. It would be to convince her that it was appropriate for a professional colleague to treat her noticeably differently than he treated his other colleagues, differently than her other professional colleagues treated her, and differently than, in her experience, professional colleagues ordinarily treat one another. Thus it would be to convince her that others were not bound by metaethical requirements of impartiality in the application of professional rules of conduct to their treatment of her, and so that she was not an equal partner in the enterprise of moral community. It is unlikely that one could rationally convince a rational moral agent that he in fact was not one; and Smith's agenda looks more like one of passive-aggressive intimidation than rational persuasion. But in the absence of any such rational assent, Washington's *de facto* cooperation with Smith's moral theory, according to which there is nothing untoward about Vogeler's treatment of her and so nothing to protest, can only be coerced – by verbal or emotional abuse, perhaps, or insinuated threats about her professional future. This is not exactly a secure basis for future moral cooperation.

So from consideration of the foregoing meta-level dilemma, we might derive at least one criterion of selection for the most adequate moral theory (or range of theories) among the alternatives:

> (1) A practically adequate moral theory *K* recognizes fully the rational agency of any full participant in the social and economic life of a community of ordinary adults, even if that person espouses a moral theory that, under particular circumstances, competes with *K* for practical implementation.

Now it may not be obvious why it is necessary to state (1) as a selection criterion, given the role of recognition used so far in the technical Kantian, cognitive sense described in Chapter II, and the role of self-recognition in a rational principle which I have already ascribed to a rationally motivated agent in Chapter V.4.5. After all, if a moral theory is a special kind of rational theory and a moral agent is a special kind of rational agent, it should go without saying that a moral agent recognizes not only himself in rational principles, but other rational agents as well, regardless of their views. Unfortunately this does not go without saying. We have already seen how our pseudorationality warps and twists the scope of our favored theories in order to satisfy our desires or assuage fears caused by anomalous threats to their rational intelligibility. This feeble self-protective mechanism obstructs our ability recognize in our moral principles those rational moral agents who, on the one hand, are deserving of such recognition; and, on the other, present challenges to our psychological capacity to extend those principles to all who in fact fall within their scope. In the example, Washington is precisely such an agent; and presents precisely such a challenge to Smith's and Vogeler's moral theories.

So it is, after all, necessary to spell out what recognition requires for the practical application of a moral theory that must be protected from our pseudorational proclivity to bend it out of shape; these ruminations apply to all four of the criteria offered in this and the next section. In this context, then, the term "recognition," implies the specifically moral and practical inflection that the term "acknowledge," often considered equivalent in meaning, has in some contexts. To *recognize* something about a person in this sense requires certain practical reinforcements to the conditions spelled out in Chapter II and thereafter, to wit:

(1.a) to acknowledge it verbally to oneself and to the person under appropriate circumstances;

(1.b) to elaborate on it verbally to oneself and to the person under appropriate circumstances;

(1.c) to facilitate verbal acknowledgment of and elaboration on it by oneself and others to the person under appropriate circumstances, such that

(1.d) these verbal declarations call up the appropriate emotions of respect and acceptance in the speakers, and motivate the appropriate behavior.

A moral theory that recognizes something about a person imposes these requirements of behavior on its proponents in the non-ideal case. That is, it requires them to express this recognition of others in their conduct toward them.

(1) requires that, in the formulation of the descriptive laws and practical principles of conduct to which a community is expected to adhere, an adequate moral theory *K* must include all recognizably moral agents – i.e. rational agents at the very least – in its scope of application, whether or not particular agents agree with *K* theoretically. It states that all deserve equitable moral treatment – and, in particular, equal recognition for their particular moral theories. It precludes drawing the lines of the community of fully moral members to which K applies such that only one's moral allies and cohorts fall within it, whereas competitors, enemies, and strangers count as morally defective outsiders.

Of (1), Postow objects that no theory can satisfy (1), "for to disagree with any rival theory is to regard as distorted some of the moral perceptions that are informed by that theory."[9] But this objection can be avoided by simply distinguishing between beliefs, principles, and perceptions; and between mistaken perceptions and distorted perceptions. If I believe you are wrong to assert that I must keep my promises in a particular instance, my reason may be simply that your general principles are too provincial. I will regard your moral perceptions as correspondingly distorted only if they are saturated by your provincial general principles; but they may not be. For example, Huck Finn's moral perception that it would be wrong to turn in Jim the runaway slave was not distorted by his provincial general principle that aiding and abetting runaway slaves was a crime.

(1) does not prescribe a single right way Smith ought to respond to Washington's allegations. For example, Smith might satisfy (1) either by engaging Washington in rational evaluation of the evidence for and implications of Vogeler's behavior, or by abdicating to the

university ombudsman the role of moral mediator on the grounds that because Vogeler is a close friend of his, he is unable to render impartial judgment on this case. But (1) does sift out pseudorational tactics of the sort Smith deploys in suggesting that Washington is "seeing things" rather than seeing clearly the intrinsically offensive character of Vogeler's behavior. (1) also rules out any moral theory that withholds full membership in the moral community from certain adult groups on the grounds that they are, by nature or by ideology, not fully competent members of that community. I dissect such theories in Chapter XI, following. Moreover, (1) rules out any Anti-Rationalist moral theory that stipulates an agent's inclusion in one's family or circle of friends or local professional network as a necessary condition for full moral treatment of her.

Finally, (1) eliminates any moral theory that justifies the devaluation or subversion of an agent's rational and evaluative faculties in order to influence his action – e.g. through coercion or manipulation. So, in particular, it eliminates Classical Utilitarianism as a viable candidate for practical moral adequacy; this conclusion underscores the argument made in Volume I, Chapter XII. Perhaps controversially, (1) implies that, in the event that the practical consequences of choosing one moral theory over another involve life and death – for example, if my rival's moral theory legitimates the killing or torture of heretics and infidels whereas mine does not, it is impermissible to deploy tactics of persuasion such as the killing or torture of my rivals, just because I anticipate their deploying those tactics against me. (1) does not exclude self-defense against one's rivals when necessary. But it does exclude any behavior that "sinks to the level" of reciprocally coercing moral assent through psychological or physical power plays against them. If the impasse between Smith's and Washington's moral theories regarding the import of Vogeler's behavior cannot be resolved without reliance on underhanded attacks on the moral and rational competence of the theorist, it cannot be genuinely resolved at all.

5.2. Recognition of Pain

(1) gives us prima facie reason to suspect Smith's moral theory because it violates (1) in its rules of conduct toward competitors for moral truth. This is damaging because it reveals that the claim to superiority of Smith's moral theory depends, not on a careful assessment of its intrinsic epistemic and practical merits; but instead on undermining Washington's status as a fully responsible moral agent. But there is more to be said about it than that, even putting aside for the moment the meta-level dilemma. Among the many things that Washington communicates to Smith is the mental and emotional anguish she feels at being the target of Vogeler's verbal assaults. Smith's response is to (a) minimize the moral importance of Washington's pain, by suggesting that her reaction is out of proportion to the events that purportedly caused it; (b) deny the causal effect of Vogeler's behavior, by suggesting that Washington's pain is largely self-generated by her tendency to see slights where none were intended; (c) dissociate Washington's pain from Smith's constellation of significant moral priorities, uppermost among which is preservation of collegial equilibrium. Let us look at each of these reactive strategies more closely.

(a) judges Washington's level of mental distress to be morally unjustified by the situation that purportedly gave rise to it. Thus it presupposes that there is some morally appropriate level of mental distress that is justified by the situation. Smith indicates what this is: It is the level of distress experienced by all untenured junior faculty members as they "run the gauntlet" of performance, evaluation, and interaction with their senior colleagues in their attempts to obtain tenure. One problem is that this inclusive criterion of justifiable mental distress is too inclusive, for it does not distinguish the kinds of professional behavior by senior colleagues that are themselves morally justifiable from those that are not. Therefore it cannot distinguish levels of mental distress in response to such behavior that junior colleagues ought to learn to take in stride from those that constitute justifiable grounds for protest.

But a larger problem with (a) is that it is circular. The idea of an appropriate, justifiable level of mental distress implies that there are some morally justified ways of treating others that can be expected to cause them a certain, justified level of mental anguish – and no more. But it is hard to imagine how this level could be specified independently of the behavior that is expected to cause it, and of who could possibly be in a position to do so. To what independent standard could we possibly appeal in order to ascertain this? No variant on the "Impartial Rational Spectator" would suffice. Suppose we could spell out the psychological and emotional makeup of some such "Emotional Rational Participant" on a statistical basis that at the same time corrected for gender, class, and ethnic bias, which is unlikely. We have already seen in Volume I, Chapter IV.1 that we still would have no means of making interpersonal comparisons among distress or happiness levels of different individuals. Therefore we would have no means of ascertaining to what extent the standard of the "Emotional Rational Participant" had been met in a particular case.

In any event, the very idea of a common standard of appropriate emotional response, independent of appropriate conduct, is suspect. No one is exempt from sensitivities on a wide range of individual and idiosyncratic matters. These sensitivities may increase the intensity of one's emotional response beyond some local convention when those sensitivities are wounded: Sensitivity to one's height or weight, to being teased or not invited to parties, to one's class background or table manners or general condition of moral dereliction are just a few of the sore spots that may elicit a more vehement response than one's audience may have expected. In these cases we do not ordinarily think such a response is inappropriate relative to some emotional norm, unless it is patently self-destructive or morally costly to others – in which case the relevant norm is not emotional but moral. Instead we are reminded of how broad and inclusive the range of acceptable emotional responses may be, and we adjust our behavior accordingly so as not to give offense in the future. Unlike criteria of rationality, which are more or less uniform and applicable across a large variety of groups,[10] emotional responses are not the kind of thing that meaningfully can be legislated across individuals. This is why Anti-Rationalist moral theories that insist on grounding moral behavior solely in some implied standard of correct moral emotion sometimes seem so arrogant. They presume to instruct us as to the sort of interior

emotional life we all ought to lead in order to enjoy moral rectitude, as though acting from conscientious and well-intentioned motives toward others were not enough.

The most serious objection to (a), then, is its moral arrogance. Smith simply is not in a position to presume knowledge of that level of mental distress that it would be morally justified for Washington to feel; and even if he were, he would have no business imposing that standard on Washington. Washington's level of mental distress may be greater than Smith is comfortable witnessing. It may be greater than Smith imagines he would feel under similar circumstances. It may even be greater than previous victims of Vogeler's aggressions have expressed to him. Smith nevertheless has no basis for claiming that Washington's reaction is excessive. Minimizing the moral importance of Washington's pain is a pseudorational tactic that excludes that pain from the domain of Smith's moral theory.

So a second criterion of selection for the most adequate moral theory among the alternatives might run as follows:

> (2) A practically adequate moral theory *K* recognizes fully the moral importance to an agent of that agent's pain, as sincerely expressed in words or behavior.

As with (1), (2) seems so obvious that, on reflection, it may be unclear why it is necessary to state it. A moral theory that prescribed disparaging, belittling or ignoring another agent's expression of pain, or was silent on the question of whether it was worth alleviating, would be no moral theory at all.

Richard Miller offers a putative counterexample in the Yanomamo,[11] but I am not convinced by his account that even the Yanomamo regard it as *morally* right to shoot their wives in the thigh for being too slow with the dinner, much less that we should accept this. Miller's defense of this thesis is based on the unquestioned extension of linguistic practices unproblematic among Yanomamo men to cases that are clearly problematic for Yanomamo wives – as though the victims of a punitive social practice should have no voice in evaluating its moral legitimacy. Moreover, Miller furnishes no substantive criterion for identifying a moral theory, or for distinguishing it from mere social or psychological conventions. Let me suggest an obvious one: A moral theory must, at the very least, provide a solution to Prisoner's Dilemma-type situations, which the Yanomamo convention of fierceness does not. For example, it decimates 25% of Yanomamo tribesmen and incapacitates Yanomamo wives from getting the dinner at all. A moment's thought will suffice to see that the point generalizes to any social convention of generally disregarding other agents' pain. It is rather for Miller to explain why we should identify a self-defeating social convention as a moral one; and why in particular the principle of disparaging, belittling or ignoring another agent's pain should enjoy moral legitimacy when no serious moral theorist would prescribe such a principle.

Yet the foregoing hypothetical case combines elements of behavior that are all too familiar in a variety of social contexts, and that are implicitly assumed to be entirely consistent with a variety of standards of moral rectitude. We often disregard or belittle another's pain, or exclude it from the domain of moral concern, or give it only cursory attention or moral weight,

simply because we disapprove of its hypothesized cause. We may judge the person to be oversensitive, or self-indulgent, or manipulative, or temperamental, or distorted in her perceptions. These are terms of evaluation that indicate that we are second-guessing the motive or causes behind the agent's expression of pain, and invoking this *ad hoc* hypothesis about the disreputable origins of that expression in character or circumstance in order to minimize its moral significance. This type of rationalization is highly vulnerable to the charge of moral arrogance just discussed. It is difficult to imagine what causal origin of pain could possibly justify taking the pain itself less seriously.

Or it may happen that an agent passes such judgment on himself. He may not realize that he is a victim of moral transgression, even though the act itself causes him intense pain, because he believes he deserves it, or that the transgressive act is unexceptionable, or that it hurts the transgressor more than it hurts him. Or he may believe about the status of his own pain any of the dismissive judgments just mentioned, if he abdicates epistemic authority about his inner states to someone else who makes them. In these cases, (2) protects the victim of moral transgression against the loss of epistemic self-confidence that often comes with being such a victim, by enjoining us to take his anguish very seriously, even if he himself does not.

It might seem that Kant's own moral theory violates (2), by subordinating sensuous empirical reactions to the dictates of the categorical imperative; so that, for example, conscience may require Washington to tell Vogeler honestly that she does not appreciate his attentions, even though she knows that this will only cause him to retaliate against her with more offensive remarks to her and about her to others, which will increase her mental distress. By requiring her to tell Vogeler the truth when that will only intensify her pain, it might be argued, Kant's moral theory subordinates the full moral importance of that pain to the impartial duty to tell the truth. But in fact Kant's moral theory has no such implication. Among its imperfect duties is the duty to render aid to one in distress, and Kant acknowledges that an agent may have occasion to fulfill this duty by rendering aid to herself – as Washington does by protesting Vogeler's behavior to Smith. Although this does not abrogate Washington's perfect duty to tell the truth, it does not require that she allow herself to be treated by Vogeler as a sitting duck, either.

(2) requires that Smith respect the moral importance of Washington's pain, but it does not prescribe a single, morally correct way he should act in order to do so. Of course there do exist moral theories that prescribe stiff-upper-lipping it in response to felt mental anguish; Stoicism might be interpreted in this manner. But at best this is enjoined in response to one's own acknowledged pain, not in response to others' pain; and not, therefore, in response to the empathetic pain one may feel in response to others' pain. Nor does (2) imply that the moral importance of an agent's pain is such that it may never be outweighed by other moral considerations. What it does imply is that it may never be ignored or belittled because of them.

5.3. Recognition of Insight

Next let's look at 5.2.(b). According to 5.2.(b), Smith denies that Vogeler actually offended Washington, by suggesting that Washington's pain is largely self-generated by her tendency to see slights where none were intended. Earlier it was suggested that Washington would have to be irrational to accept the suggestion that Vogeler's intrusive and personal remarks to her, and his disparaging comments to others about her, were anything less than obviously offensive. Yet it is possible that, as Smith maintains, Vogeler's behavior was nevertheless not the main cause of Washington's pain. And it is also possible that Washington wrongly imputes offensive intent where none exists.

To see this more clearly, consider an analogous case, that of the insensitive busybody. Once the insensitive busybody finds out that you have failed your law boards or are getting a divorce, you will never be allowed to forget it. In his concern for your distress, the insensitive busybody never fails to ask you how you are handling the disappointment, and to express concern for your wellbeing and state of mind. Whenever you encounter the insensitive busybody socially, he will dilate upon this topic at length: will commiserate, suggest coping strategies, recommend relevant readings, and solicit the opinion of others as to how you should best manage your personal crisis. At first you may be gratified by his concern. But after awhile, it will be difficult not to take offense at his continually dwelling on your professional or social inadequacies. And it will be difficult not to suspect that he intends to remind you of those inadequacies, even if in fact he has no such intention. If he has none, it will be true both that he is not the sole cause of your pain, and also that you are imputing offensive intent where none exists. For at this point the other, and perhaps main cause of your pain will be your false imputation to the insensitive busybody of the offensive intent to remind you of your inadequacies. It will be your mistaken assumption that he intends to cause you pain that causes you pain, more than anything he actually does.

It is possible that Vogeler is like the insensitive busybody: tactless, insensitive, frightened, insecure, lacking both in social skills and in the modal imagination necessary to envision the psychological effect of his behavior on others – but nevertheless guileless. It may be, in short, that Vogeler is a basket case; and that the diplomatic response would be to ignore him, as Smith suggests. But even if this explanation of Vogeler's behavior were accurate, it would not acquit him of causal responsibility for Washington's pain. That pain is caused, not only by her putative tendency to see offensive intent where none exists, but also by Vogeler's deliberate *behavior*, which is intrinsically offensive regardless of intent. Nor would this explanation of Vogeler's behavior acquit him of moral responsibility for Washington's pain: if he is not enough of a basket case to be relieved of his professional responsibilities as a senior colleague, he is not enough of a basket case to be excused for not fulfilling them, either.

Moreover, Smith wrongly implies that his greater familiarity with Vogeler's personal foibles furnishes a more adequate information base upon which to evaluate the moral significance of Vogeler's behavior: Having known him from college, Smith claims, he knows

better than to interpret Vogeler's behavior as morally blameworthy. But Smith's greater knowledge of Vogeler does not necessarily translate into a more informed moral evaluation of him. It may be that, although Washington hardly knows Vogeler personally at all, she has often encountered individuals like him in the past. It may even be that she hardly knew any of them personally either; yet she still may be in a position to make a more informed moral evaluation of Vogeler than Smith. For it may be that racists and misogynists almost always are basket cases in precisely the way Vogeler is; that they never mean any real harm, but are instead reacting only to their own inner anxieties, nightmares, and resentments, without the modal imagination or sensitivity to envision the psychological effect of their behavior on others. But it is hard to see why their primitively egocentric brutality should be thought to abrogate their accountability for those effects. Washington may have no interest in speculating on Vogeler's intentional states, nor consider those states relevant to the question of whether or not his behavior constitutes harassment or not. For the primary features of Vogeler's behavior relevant to Washington's moral interpretation of it are its disparity relative to publicly affirmed norms of collegial professional conduct, and the corrupt system of personal values Vogeler reveals to Washington by engaging in it.

Thus Smith cannot argue that his special access to Vogeler's intentional states, which Washington lacks, furnishes him with an information base for evaluating Vogeler's behavior that is superior to Washington's. It may be that Washington's extensive past experience with this kind of behavior more than outweighs any insight she may lack into its phenomenal causes in this particular case. Moreover, it would be difficult to overestimate the importance and quality of the insight Washington gains into Vogeler's moral character solely from her special access, which she shares with no one else, to his racist and misogynist proclivities. By being their object, Washington thereby discovers in Vogeler morally significant character dispositions with which Smith may be unfamiliar, and that cannot be overridden by what Smith does know about him. Of course these proclivities may coexist with being a wonderful colleague and memorable school chum to Smith. But these positive qualities hardly can be invoked as a justification for denying the existence of the more dangerous ones as well. This would be as irrational as invoking Vogeler's racist and misogynist behavior to Washington as evidence that he was incapable of being a wonderful colleague and memorable school chum to Smith.

A third criterion of adequacy for a practicable moral theory might therefore run as follows:

> (3) A practically adequate moral theory *K* recognizes fully the moral importance of the insight into an agent's character a patient gains as the recipient of the type of act in question.

(3) can work to the benefit of the agent as well as to that of the patient of the blameworthy action. To extend the example: Jones, a cantankerous, foulmouthed, misanthropic senior colleague of Washington's, known far and wide for his vitriol against all things politically correct, may surprise Washington and everyone else by taking her part, mentoring her, encouraging her work,

or by resolving the issue under moral mediation that Smith may feel too implicated to address. For of course there is no inherent incompatibility between being cantankerous, foulmouthed and misanthropic on the one hand, and fair, supportive and impartial on the other. To be sure: in such a case, Washington, or any such recipient of Jones' beneficent actions, would have to do a fair amount of higher-order theorizing about Jones' true character, in order to square those initiatives with his misanthropic public behavior. Indeed, this is the sort of superficially anomalous behavior that should stimulate higher-order theorizing among inquiring minds. The outcome of that intellectual labor would be special insight into Jones' character that only someone who, like Washington, had experienced both sides could obtain.

But in the example as originally presented, in which Washington uncovers the rotten underside of an angelic public persona rather than vice versa, (3) blocks the pseudorational tactic of denying the facts of moral responsibility by denying the epistemic validity of the victim's knowledge of the transgressor. Hence just as (2) safeguards the moral importance of the pain a victim suffers at the hands of her transgressor, (3) safeguards the moral importance of the information about the transgressor a victim obtains at the hands of that transgressor. Just as we are sometimes tempted to discount a victim's pain because we devalue its circumstances of origin, so are we similarly tempted to discount a victim's perception of wrongdoing because we devalue her status as a victim, or her social relation to the transgressor, or to the system of social practices that may bestow legitimacy and status on that transgressor.

So, for example, a woman who suffers physical abuse at her husband's hands must battle the skepticism and resistance of law enforcement agencies governed by men, most of whom are also husbands, and some of whom also abuse their wives. An African American who suffers employment discrimination at the hands of a European American employer who, because of segregation is insensitive to issues of racial equity, must then battle the skepticism and resistance of regulatory agencies staffed primarily by European Americans who, also socialized in segregated environments, are equally insensitive to issues of racial equity. Or a homosexual who suffers harassment at the hands of delinquent teenagers must battle the skepticism and resistance of a largely heterosexual public. (3) protects the moral significance of the victim's special insight into injustice even when (or particularly when) the preponderance of social practices and the weight of collective skepticism are allied against her.

These two devaluations – of a victim's pain and of a victim's insight into the transgressor – are not unrelated. When an agent commits a moral transgression from a position of credibility and authority, part of what constitutes that position of power surely must be *em*powerment, in the form of the presumption of moral rectitude, by the same community that confers legitimacy and status on that agent in the first place. So it is unsurprising that members of that community might be reluctant to withdraw that presumption by giving a privileged place to accusations which, if well-founded, would have precisely that consequence; and unsurprising that it might deny equal empowerment, legitimacy and status to the accuser. Moreover, we have seen in the preceding chapter that preserving one's view of an acquaintance or colleague as a paragon of

moral rectitude is a natural expression of a more general form of pseudorationality. Vigilant self-defense is needed against the loss of moral innocence threatened by the clear and unvarnished presence of moral corruption, for it sullies those who witness it. The attraction of denying, dissociating, or rationalizing away the bad news that the victim has to disseminate is evident.

Thus (3) is needed in order to balance a natural tendency to assume that tempting viewpoint on the moral interpretation of action discussed in Volume I, Chapter XII, namely the viewpoint of the *cognoscenti* of one's favored moral theory. This is that self-defined subgroup that not only knows and avows the theory in question, but also implicitly regards itself and its members as embodying the theory's ideal of moral rectitude. Although virtually any moral theory may generate a *cognoscenti* among its proponents – the Bloomsbury devotees of Moore's Ideal Utilitarianism being a particularly obnoxious example of this, some moral theories are more susceptible to this form of corruption than others. Moral theories that stipulate as a condition of moral knowledge a special faculty or insight that not all members of the moral community can have are particularly vulnerable to this form of abuse because they implicitly arrogate possession of the special quality to the moral theorist, and invite the inference that one's special faculty or insight sanctify one's behavior as morally acceptable even if it diverges sharply and noticeably from the plebian, Golden Rule brand of moral conduct by which most of us feel obligated. These *cognoscenti moral theories* that stipulate an esoteric inner circle possessing special moral wisdom that ordinary moral agents lack, and by which even the moral victims among them must be guided, include Classical Intuitionism, understood as the view that we discover what to do by consulting a special, mysterious moral faculty which not everyone may have;[12] Classical Utilitarianism as propounded by Sidgwick, according to which knowledgeable Utilitarians are obligated by a set of moral rules different from and superior to those that enjoin the common run of people;[13] and those brands of Marxism that ascribe special, revolutionary knowledge either to the intelligentsia or to the proletariat, in accordance with whose dictates the classless society is to be realized. *Cognoscenti* moral theories violate the criterion of inclusiveness by denying to some moral agents the epistemic authority and credibility necessary for contributing substantively to moral consensus, while supplying it to others. They thus obstruct the moral agency of those so deprived, and encourage abuses of power by those thereby empowered.

(3) rules out such *cognoscenti* moral theories because they implicitly presume that membership in the relevant *cognoscenti* involves the highest condition of moral knowledge – superior, in particular, to that any nonmember moral victim might gain from being the recipient of moral vice. Unlike a Kantian moral theory, which supplies metaethical principles of derivation from which commonsense moral precepts available to all and compatible with many such theories can be derived, *cognoscenti* moral theories implicitly presume a connection between moral rectitude and epistemic familiarity with those theories themselves. Because devaluation of a nonmember victim's knowledge of moral transgression relative to a member's is built into the very structure of these theories, they violate the criterion of inclusiveness.

Of course, like any practical principle, (3) may be abused, by constructing a *cognoscenti* of moral victims. Theories that ascribe a privileged status to suffering, as some forms of Christianity do, may be particularly susceptible to this. Nevertheless (3) does provide a counterweight to the empirically more prevalent impulse to discount as false, mistaken, or misguided the insights into moral character to be gained through being on the receiving end of moral vice. It would be consistent with conformity to (3) for Smith to weigh Vogeler's collegiality and shared history with Smith more heavily than his moral turpitude, and more heavily than Washington would in deciding what should be done about it. But it would not be consistent with (3) to deny the legitimacy of Washington's insights into Vogeler's character altogether.

6. Nonrecognition of Bully Systems

Finally, consider 5.2.(c), which dissociates Washington's pain as unimportant relative to Smith's constellation of significant moral priorities, one of which is to preserve collegial equilibrium. This response not only ranks maintaining the collegial *status quo* more highly than alleviating Washington's emotional distress. It ranks more highly a *status quo* that licenses Vogeler's unjustifiably inflicting pain on Washington. On the face of it, it certainly would seem morally unjustifiable to discount the mental distress of a moral agent for the sake of preserving in equilibrium a social and professional network that deliberately and unjustifiably inflicts such emotional harm. But there are moral costs involved in reforming it that must be figured into the equation. Is alleviating Washington's pain worth the pain, inconvenience and disturbance it would cause Smith, Vogeler, and others in the department to change the *status quo* and reform their behavior? Is it worth the resentments, embarrassments, incriminating revelations, betrayed loyalties, ruined friendships, and destroyed professional equilibrium that now exists? And what about the daily work of running the department, tasks ably discharged by the very same individuals who fill Washington's life with undeserved misery?

Millian liberalism might formulate this issue as one of whether the rule of the majority or the rights of individuals should prevail, and there is much to be said for such an analysis.[14] But examination of the social relationships that knit the majority together as a majority in this case suggests an alternative formulation. The issue can also be framed as a crucial point of opposition between Rationalist moral theories and Anti-Rationalist views that postulate the priority of personal loyalties and emotional attachments over impartial duties to others. In Chapter I, I argued that this conflict – essentially a conflict between rationality and power – is most centrally definitive of the two alternatives with which the professional practice of philosophy itself is confronted. On this analysis, the fundamental question is whether it is worth unraveling an entire network of personal and professional attachments in order to rectify the injustice done to a single, unassimilated individual.

To this question, Anti-Rationalist claims about the importance of sympathy, caring, friendship, and so forth can provide no satisfactory answer, since these are the relational attributes that, in the case at hand, generate the problem. Of course an Anti-Rationalist might

just bluntly disavow the importance of Washington's anguish when compared to that which would be incurred by shifting the *status quo* in order to ameliorate it. Alternately, the Anti-Rationalist might solve the dilemma by assigning a higher priority to whatever personal or professional attachments he may have to Washington. However, to weight these relational attributes in this instance in favor of Washington is to betray precisely that network of personal and social ties on the importance of which an Anti-Rationalist moral view insists. At best, an Anti-Rationalist might plead divided loyalties in this case. But because an Anti-Rationalist moral view admits of no strictly impartial principles above and beyond the spontaneous dispositions of character that motivate individual interactions, it can furnish no higher-level principles for adjudication between such conflicting loyalties.

By contrast, a rationalist moral theory tackles the solution to this problem quite straightforwardly, because it includes all fully functioning moral agents within its domain of explanation. And in virtue of its aspiration to legitimacy as a genuine theory, it emphasizes satisfaction of the metaethical requirement of impartiality in the application of its laws and principles. Thus as we have seen, a rationalist moral theory rules out violations of (1), above, on impartialist grounds, because these fail to treat a moral agent as an equal member of the moral community. But 5.2.(c), above, violates (1), because it implies that since Washington is an interloper in and potential disrupter of the collegial social network rather than a fully integrated member of it, she is unentitled to full moral treatment by its members. For a rationalist moral theory, this is unacceptable.

Secondly, 5.2.(c) violates (2) because it dissociates Washington's pain from the domain of moral importance in which Smith situates the pain Vogeler would feel at being reprimanded for inflicting it, the pain Smith would feel at having to reprimand him, and the pain both would feel at the way this episode undermined preservation of their professional connections more generally. But surely Washington's pain is not outside the moral domain of Vogeler's or Smith's. Surely Washington's pain is to be weighed in the same balance with Vogeler's and Smith's, and, because Washington's pain is an unjustified moral harm whereas Vogeler's and Smith's pain would be the result of a justified moral restitution, to be found of greater moral weight than both. This suggests that Smith's and Vogeler's pain is morally permissible as a means of alleviating Washington's morally impermissible pain. Smith's dissociation of Washington's pain from the domain of moral significance is a pseudorational attempt to protect his social network at the expense of social justice.

A moral theory that assigns greater value to preserving a system's stability than it does to alleviating unjustified pain in a particular case is thinkable, even if the primary purpose of the system is to alleviate pain so far as is possible, and may even be warranted under some circumstances. But a view of the sort expressed in 5.2.(c), which assigns greater value to the preservation of a system whose very stability depends on permitting the infliction of unjustified harm – call this a *bully system* – is not. International examples of bully systems include Hitler's Germany, Ceausescu's Romania, Botha's South Africa, Milosevic's Yugoslavia, Taylor's Liberia,

and, of course, the United States throughout its short but brutal history. A bully system legitimizes harm to moral victims as a necessary means to the preservation of equilibrium – and the benefits of equilibrium – among moral transgressors. It condones the protection of moral transgressors from the punitive consequences of their transgressions; spreads the benefits of that protection among all such transgressors; and concocts a pseudorational ideology that simultaneously rationalizes those benefits and denies or dissociates the rights of the transgressed whose pain pays for them. In essence, a bully system comprises a community of *übermenschlichen* free riders who, when it promotes self-interest, violate their social covenant with their *Untertanen* victims. This is a particularly cynical travesty of moral principle.

In general, a moral theory that aspires to conform to the metaethical requirement of impartiality cannot condone social practices that even occasionally permit harm to the innocent in order to evade punishment and accrue benefits for the guilty, on pain of perverting the meaning of the words "innocent" and "guilty". By treating the innocent as guilty and the guilty as innocent in those cases in which the moral victim is seen as outside the social network, bully system practices make impossible the consistent application of punitive sanctions to all those ostensibly picked out by a rationalist moral principle. And by thus violating the requirement of impartiality, they thereby, in this case, violate that of inclusiveness as well.

We may attempt to capture this conclusion as follows:

> (4) A practically adequate moral theory *K* assigns greater weight to protecting an agent from harm than it does to protecting a bully system from the punitive consequences of harming her.

(4) ensures that the moral laws that govern a network of moral agents are not distorted or tailored so as to effectively legitimize harmful behavior by its members. Although it does not provide a specific answer to the question of how best to rectify the harm done to Washington by Vogeler, it does ensure that preservation of a morally corrupt network does not become an end in itself, to which the value of morality itself is subordinated. And it stipulates that in a run-off between rectifying injustice to an individual and preserving unjust practices that stabilize and benefit a group, the former will take clear precedence over the latter. This means that (4) rules out Anti-Rationalism as a valid moral theory, since it permits the opposite order of precedence in some cases. (4) thus elaborates the criterion of inclusiveness to cover those situations in which, although an agent may be acknowledged by the group as an agent and his pain ascribed full moral importance, his agency and his legitimate demands for assistance or restitution are not considered sufficiently weighty to take precedence over preserving intact the corrupt but stabilizing practices that cause that pain.

Earlier it was suggested that there do exist moral considerations that might reasonably outweigh the *prima facie* duty to relieve an innocent agent's suffering; but preserving a bully system's equilibrium by permitting its members to inflict such suffering is not one of them. By constraining the application of moral principles of aid or restitution only to members of the group or network, or perverting their application so as to relieve moral transgressors of

accountability, a bully system both narrows the scope of application of the theory and manipulates the formulation of its principles so as to exclude outsiders from its full protection. (4) redresses that exclusion, by withholding moral recognition from such a system; and, in particular, by withholding verbal acknowledgment, and elaboration and facilitation of such verbal acknowledgment, that might call up emotions of respect and acceptance of it in the speakers, and motivate behavior expressive of these emotions (Section 5.1, (1.a) – (1.d), above). Regardless of the advantages or attractions a bully system may offer, it deserves neither our respect nor our acceptance but rather our condemnation.

7. "Seeing Things"

(1)-(4) obviously have many other applications beyond those examined in the hypothetical case I have invoked to derive them. And it is unlikely that (1)-(4) constitute the only criteria of inclusiveness a practically adequate moral theory must satisfy. But I would maintain that they at least constitute a significant subclass of them, because each responds to a familiar, pseudorational strategy by which relevant moral data are typically excluded from moral consideration.

Among the main contenders for practical adequacy I have examined, a Kantian-*type* moral theory appears to be the only one capable of satisfying each of (1)-(4).[15] To review the arguments of this chapter: Classical Utilitarianism licenses less than full acknowledgment of a person's rational agency when this promotes general welfare, thus violating (1). Anti-Rationalism does the same when the agent in question is not personally attached to the right social network. Classical Utilitarianism, Intuitionism, and certain varieties of Marxism devalue a victim's moral knowledge relative to that of any arbitrarily selected *cognoscenta*, thus violating (3). And both Anti-Rationalism and Classical Utilitarianism permit the devaluation of a victim's claim to aid or restitution when this threatens a bully system's stability and personal attachments, thus violating (4). Only some variant on a Kantian theory seems able to resolve satisfactorily the initial dilemma of moral interpretation with which this discussion began, because only a Kantian-type theory unambiguously includes all the data of Washington's predicament within the moral domain, and respects fully their importance once there. That Washington's interpretation of Vogeler's behavior as harassment is accurate has been clear from the outset. That Washington's interpretation presupposes a Kantian-type moral theory that satisfies these criteria of inclusiveness, whereas Smith's interpretation does not, may help explain why Washington is not just "seeing things," as Smith maintains, but rather is seeing things considerably more clearly than he.

Endnotes to Chapter X

[1]In the discussion of moral theory that follows, I reserve use of the term "laws" to refer to the components of ideal descriptive, explanatory moral theories, and "principles" to refer to their prescriptive practical applications for imperfect human beings.

[2] "Piper's Criteria of Theory Selection," *Southern Journal of Philosophy XXIX, Supplementary Volume: Moral Epistemology* (1990), 60 – 65.

[3] Harman, Sturgeon, Boyd.

[4]Thomas Kuhn, *The Structure of Scientific Revolutions* (Chicago: University of Chicago Press, 1971), Chapters VI-VIII.

[5] Surprisingly, Kuhn relegates the criterion of inclusiveness to a second-class status, along with simplicity and compatibility with other specialties (*ibid*., Postscript: 206). I find this surprising since the more inclusive in scope a theory is, the less vulnerable it is to the de-stabilizing effect of anomaly.

[6]Peter Singer, *Animal Liberation*, Second Edition (New York, NY: New York Review Books, 1990).

[7]Of course the distinction between theoretical terms and observational terms can be ultimately only a matter of degree, rather than of kind, to the extent that it is valid at all. See Norwood Hanson, "Observation," in Richard Grandy, Ed. *Theories and Observation in Science* (Englewood, N.J.: Prentice-Hall, 1973), 129-146.

[8]An easy way to keep clear the cast of characters is to connect Vogeler's name with the double entendre in the German vernacular.

[9]*Op. cit.* Note 2.

[10]In "African Traditional Thought and Western Science," (in Bryan Wilson, Ed. *Rationality* (New York: Harper and Row, 1970), 131-171), Robin Horton furnishes convincing evidence for the cross-cultural validity of at least some fundamental norms of theoretical rationality.

[11] Richard Miller, "Ways of Moral Learning," *The Philosophical Review XCIV*, 4 (October 1985), 507-556). More recently, questions have been raised about the fidelity of canonical anthropological accounts of the Yanomamo. But this does not affect their philosophical interest, even if we must relegate these accounts to the hypothetical or fantastic.

[12]Sir David Ross develops this idea in *The Right and the Good* (Oxford: Clarendon Press, 1973), 29-33.

[13]Cf. Henry Sidgwick, *The Methods of Ethics* (New York: Dover Publications, 1966), Book 4, Chapter 5, Section 3. I examine Sidgwick's view in Volume I, Chapter XII.

[14]Cf. Ronald Dworkin, *Taking Rights Seriously* (Cambridge, Mass.: Harvard University Press, 1977), Chapter 7.

[15] I have not examined so-called "virtue theories" that purport to be based in Aristotle's ethics because I do not think this particular appeal to authority is well-grounded. As we know, Kant also has a virtue theory that he develops in MM Part II; and Aristotle arguably gives the

same pride of place to reason and rational deliberation as does Kant. No viable "virtue theory" can be coherently articulated without reference to guiding rationality principles of the sort that both Kant's and Aristotle's normative moral theories contain, and that provide the central focus of attention in this discussion. For further discussion of Aristotle's own moral theory, see Volume I, Chapter V.1.4. This original in any case falls short of satisfying (1) – (4), because it straightforwardly fails the general criterion of inclusiveness in the ways indicated in Section 2.3 and 2.5 above.

Chapter XI. Xenophobia and Moral Anomaly

Chapter X elaborated some criteria of inclusiveness fashioned specifically for the non-ideal reality of a practically adequate moral theory. These criteria were intended in part to redress the harmful consequences of pseudorationalizing significant moral phenomena out of the realm of moral concern – i.e. of turning such phenomena into first- or third-person moral anomaly that functions as an enigmatic and alien threat that therefore undermines and destabilizes the theory. I suggested that these criteria themselves had moral import, in that they counteracted the exclusion of morally significant agents, actions, events or states of affairs from the realm of moral concern; and therefore counteracted the pseudorational dehumanization and demeaning treatment of morally significant agents as third-personal moral anomaly. Satisfaction of these criteria of inclusiveness would not ensure the moral rectitude of all recognized members of the resulting moral community. But it would ensure that no moral agent were viewed as so much of an enigmatic and alien threat to one's favored moral theory, to one's honorific self-conception and therefore to the interior coherence of one's self that such an agent's capacity for rational agency and therefore moral accountability themselves were denied. So, in particular, satisfaction of the proposed criteria of inclusiveness would tend to defeat the disposition to xenophobia.

The xenophobic response marks the outermost boundary of our pseudorational response to third-personal moral anomaly more generally. In Chapters VIII and IX I considered examples in which another person's behavior violated our favored moral theory in ways that impelled us merely to rationalize it. More serious and disruptive violations were met with dissociation. The occurrence of third-personal moral anomalies that were even more morally unacceptable were simply denied out of existence. I did not attempt to correlate the seriousness of the violation with the magnitude of moral harm done, since I do not believe there to be any such correlation. However, there are two cases in which we exclude third-personal moral anomaly from membership in a moral community that do correlate with magnitude of moral harm. The first is our instinctive reaction to a moral agent's deliberately infliction of unthinkable moral evil; Hitler would be the most familiar but not the only instance of such an agent. I suggested in Chapter IX that we react to such instances as assaults on the very concept of morality our moral theory expresses; i.e. we assign Hitler and the like the status of incorrigible moral anomaly. But the second case is even more radical than this, for in essence it mirrors what Hitler and others like him actually do. By deliberately disregarding, diminishing or destroying others' rational agency – i.e. by dehumanizing them, Hitler and others like him exile recognizably rational agents from the moral community of such agents altogether. This is the attitude toward and treatment of others that defines xenophobia, and the treatment against which Chapter X's criteria of inclusiveness were proposed as an antidote.

In this concluding chapter I dissect in detail in what the disposition to xenophobia consists, in order to understand better how and why a practically inadequate moral theory can

fail to satisfy these criteria of inclusiveness. This analysis does not purport to be exhaustive, or to supply necessary and sufficient conditions of practical inadequacy for a moral theory. But it does aim to systematize some of the more familiar and recognizable breaches of inclusiveness that give so much of our actual moral conduct its peculiarly clubby flavor. My approach assumes that the xenophobic response is an innate, hard-wired, inescapable part of our cognitive make up that we cannot eradicate; but also that it is possible to control it and guide its targets appropriately through social conditioning. I also assume that xenophobia is not best understood as a transaction between different groups, but rather as a transaction between individuals in interpersonal relationships. Indeed, the most pressing question a competent analysis of xenophobia must answer is how such abstractions as nation, race, ethnicity, or religion can turn neighbors, friends, couples, colleagues or co-workers into enemies virtually overnight; I address this question directly in Section 5, below. Thus individual transactions have important implications for different racial, ethnic, or social groups and the interactions among them. But on the proposed analysis, xenophobia initially and primarily occurs in transactions between individuals: partners or friends or relatives or co-workers or neighbors or fellow citizens. If we are to understand the behavior of larger groups, and of the official representatives or delegates of these groups, we need to understand these more elemental interactions first. So my analysis presupposes methodological individualism.

In Chapter II I argued that our scope of judgment is confined to those properties and particulars that conform to pre-existing categories and concepts that structure not only our experience, but thereby our selves. I also argued that we are compelled either to conceptualize the objects of our experience in familiar terms, or else not to register them at all; and that this is a necessary condition of preserving the unity and internal coherence of the self against anomalous data that threaten it. Correspondingly, in Chapter VII I argued that resistance to integrating conceptual anomaly is a general feature of human intellection that attempts to satisfy the Kantian requirement of literal self-preservation described in Chapter V.2. I locate my analysis of xenophobia within this context. I invoke the proposed Kantian conception of the self to explain the phenomenon of xenophobia as fear of another who fails to satisfy our provincial preconceptions about bona fide persons; and xenophobia, in turn, to explain the phenomenon of political discrimination.

Relative to this broader Kantian conception, xenophobia is – like scientific resistance to natural phenomena untamed by theory – a special case of a perfectly general disposition to defend the self against anomalous informational assaults on its internal coherence, i.e. the highest-order disposition to literal self-preservation. Thomas Kuhn documents the form this resistance can take in the natural sciences: the inherent impediments to paradigm shift, the conservatism and constitutional insensitivity to the significance of new data, and the resistance to revising deeply entrenched theories in light of experimental anomaly that are all by-products of scientific method and professional practice.[1] In a similar manner, an agent who is personally invested in a provincial moral theory views as morally anomalous another agent who violates her

correspondingly provincial conception of how moral agents should behave or appear, and reacts to that violation xenophobically. So on the analysis I offer here, xenophobia is fear, not of strangers generally, but rather of a certain kind of stranger, namely the kind who does not conform to one's preconceptions about how persons are supposed to look or behave. It is a response to a very specific kind of third-personal moral anomaly: not the kind that violates a principle of right conduct a moral theory prescribes; but rather the kind that violates unspoken empirical preconceptions about the kind of agent who is legitimately held to that principle in the first place. Xenophobia is first and foremost a reaction to self-generated appearances rather than to independent realities.

Section 1 sketches an alternative analysis of xenophobia, which I call the Marxist analysis, against which I contrast and highlight my own. Section 2 gives an overview of that Kantian analysis of xenophobia, and situates it within the analysis of pseudorationality offered in Chapters VII and VIII. Section 3 bears down on the details of this analysis with regard to its precedent in Kant, and describes three constitutive errors of cognitive discrimination, in terms appropriated more or less directly from *The Critique of Pure Reason*. Section 4 offers the fourth and final test case for my analysis of pseudorationality, namely political discrimination, and a detailed analysis of its functioning. Section 5 derives from the Kantian conception of the self developed in earlier chapters two ways in which failures of cognitive discrimination might function, the second of which is consistent with its corrigibility. Section 6 makes the case that Kant's own analysis of reason favors this second alternative, and implies an antidote to xenophobia in xenophilic curiosity about and interest in the unfamiliar. Section 7 discusses the arena of contemporary art as a training ground for cognitive discrimination in which the xenophilic impulse might be cultivated. Finally, Section 8 makes explicit the implications of my methodological individualism and Kantian rationalism: that the essence of xenophobia is to be found in all interpersonal relationships – alongside the capacity for moral alienation and moral heroism; and that all of these are by-products of the challenge to make rationally intelligible inherently enigmatic concrete particulars.

1. The Marxist Analysis of Xenophobia

We commonsensically associate xenophobia with racial or ethnic hatred. Various examples of xenophobia that support this association include racism, misogyny, homophobia, anti-Semitism, and ethnic hatreds of various kinds. These are examples of generalized revulsion directed against other moral agents fully or partially excluded as theoretically anomalous from the scope of our favored moral theories. However, I adhere to the more literal meaning of the term, which is fear of strangers. We tend to hate – or, at the very least, strongly dislike – what we fear. But fear is the more primitive response to geographic, physical, or psychological threats to the integrity of the self. So this literal definition implicitly supports my thesis that xenophobia is merely a special case of a more general pseudorational response to any kind of anomaly that threatens the interior coherence and rational intelligibility of the self.

As a foil to my account, I sketch in this section what aims to be a plausible version of a contrasting view one might hold. On this Marxist analysis of xenophobia, we fear the stranger in question, not because he threatens our sense of ourselves as coherent and rationally intelligible agents; but rather because we anticipate that he is going to deplete the economic and material resources of our community. We assume that there are not enough of these resources to go around, and that he will consume more than his share; "his share" being proportionally reduced in accordance with his alien status relative to our moral community. Since, we reason, he is not really one of us, he is not entitled to consume as much as each of us who is. If he does anyway, we view him as presumptuous and transgressive; and instinctively close ranks in order to protect ourselves and our resources against his further incursions.

This type of analysis has a certain intuitive plausibility, but also raises several questions. First, when we instinctively assign to another a smaller measure of resources than those we assign to ourselves and our friends, we already thereby signal our exclusion of the other from full membership in our moral community, and therefore our xenophobic reaction toward her. Because the Marxist analysis begins with the assumption of a prior, inequitable assignment of resources, it effectively begs the question an analysis of xenophobia should answer, namely who fails to get what and why. The Marxist analysis does not explain who counts as a member of our community who is deserving of a fully apportioned share of economic resources; or how one ever comes to make the distinction between ourselves and the members of our community with whom we identify on the one hand; and, on the other, the alien outsider whom we regard as different. How do we decide whom to close ranks against? Is the criterion looking different, or speaking with an unfamiliar accent? Or is it having a different personal history or sexual preference or values that are not visible in one's appearance or self-presentation? Or is it sharing a common history and physical appearance but then feeling betrayed by someone who is discovered to have different values, preferences, or racial or ethnic affiliations? Or by someone who in fact has the same values, preferences, or racial or ethnic affiliations we do, but looks and acts completely differently? Whereas my analysis addresses these questions, the Marxist analysis begs the question as to how to identify the object of xenophobia, by presupposing an answer to the question of who is not one.

A second question that the Marxist analysis does not address is this. If the basic issue is the distribution of economic resources and the scarcity thereof, why should we be so loyal to our own community, however delineated, and to what extent are we in fact? Scarcity of economic resources exists among those of us within the community as well as between our community and those outside it. So why should each of us not close ranks against everybody else, rather than just the strangers we designate as different? Whereas my analysis acknowledges and accounts for this possibility, the Marxist analysis seems to entertain sentimental fantasies about group solidarity that conceal the ease with which actual communities can disintegrate into roaming hordes of free riders.

The Marxist analysis of xenophobia as I have briefly sketched it also has some infelicitous implications. First, it seems to imply that xenophobia should decrease in direct proportion to the threat to the community's economic security. That is, the less of an economic threat the stranger is, the less he should be the object of xenophobia because the less anyone has to fear from him economically. But non-ideal reality does not confirm this implication. For example, consider American racism. Less economically competitive African Americans are more feared and hated than the relatively well-off, middle-class African Americans who really do compete with European Americans for economic resources. Those trapped in ghettos, without education or marketable skills or life prospects, those unable to compete for jobs, are seen as more threatening. They are more readily and frequently demonized in the media and popular culture than those who have successfully assimilated into the middle class; and for this reason are more readily treated as xenophobic objects.

Now the Marxist may retort that this example in fact supports the Marxist analysis, because, first, it demonstrates that middle-class African Americans are fully assimilated not only into the middle class, but thereby into the moral community more generally; and therefore are neither regarded as alien outsiders nor treated as xenophobic objects. However, middle-class African Americans know that this reasoning is false. Second, the Marxist may argue that it stands to reason that those who are more economically deprived and consequently have less to lose are correspondingly angrier, more desperate, harder to control and therefore more of a threat to those classes whose wealth they view and covet from a distance. This may well be true. But members of the underclass are not an economic threat, nor regarded as economically competitive with the well-off for scarce resources. Thus this argument contradicts the implication of the Marxist view, that those outsiders whom we anticipate will consume more of the community's economic resources are seen as more of a threat, and those whom we anticipate will consume less are seen as less of one. On the contrary: the Marxist's retort implies that those outsiders who consume less of the community's resources are therefore seen as more of a threat. If this is true, then the community's fear of strangers cannot be unpacked as a fear that they will merely consume more than their comparatively reduced share and consequently deplete the community's economic resources. The real fear is that those who consume less eventually will consume the entire community in a suicidal conflagration of rage, violence and despair. This fear is not merely, or even primarily, about the loss of economic resources alone.

The Marxist analysis of xenophobia also has the implication that, in so far as the worry really is about economic resources and nothing else, then the more economically secure a community is, the less xenophobic it should be. That is, the more resources it has, the more resources it has for protecting those resources, and the more resources it has for generating those resources. All these resources should increase the community's sense of security, and thereby make it at least somewhat more receptive to outsiders. This, too, is exactly the opposite of what we find in reality. Again consider American racism. Here we find that it is the more economically secure who seek segregated neighborhoods, workplaces, housing, gated

communities, guards, doormen, guns, fences, walls, security systems, offshore bank accounts, and social contact exclusively among those of similar class, racial, ethnic and religious background. Indeed, the importance of these self-insulating measures seems to increase rather than decrease with level of economic well-being. It would appear, then, that the more economically secure a community is, the more ingrown, isolated, self-protective and hostile to outsiders it becomes.

Again the Marxist may retort that this case also supports rather than undermines the analysis; for it stands to reason that those who have more economic resources have more to protect and more to lose, feel therefore more vulnerable and insecure and so must be correspondingly more vigilant against intruders. This, too, may well be true. But it, too, directly contradicts the implication of the Marxist analysis, that increased economic security is correlated with decreased xenophobia. If it is true that those who have the most economic resources are the most threatened by those outsiders who have the least, then no quantity of accumulation or protection of economic resources is sufficient to quell that fear, for an equal redistribution of them to formerly deprived outsiders might well be anticipated only to exacerbate the stringency of their demand for restitution. In that case, the existence and degree of fear of the alien other is independent of the existence and degree of a community's economic accumulation. The fear that characterizes xenophobia has a different cause.

2. A Kantian Analysis of Xenophobia

I shall use the terms *person* and *personality* to denote particular empirical instantiations of the concept of personhood, which I assume to be innate for purposes of this discussion. Thus when we refer to someone as a person, we ordinarily mean to denote at the very least a social being whom we presume – as Kant did – to have consciousness, thought, rationality, and agency. The term "person" used in this way also finds its way into jurisprudence, where we conceive of a person as a rational individual who can be held legally and morally accountable for her actions. Relative to these related usages, an individual who lacks to a significant degree the capacities to reason, plan for the future, detect causal and logical relations among events, or control action according to principles applied more or less consistently from one occasion to the next – i.e. who lacks the capacity for resolute choice in McClennen's sense – is ascribed diminished responsibility for her actions, and her social and legal status as a person is diminished accordingly.

Similarly, when we call someone a "bad person," we communicate a cluster of evaluations that include, for example, assessing his conscious motives as corrupt or untrustworthy, his rationality as deployed for maleficent ends, and his actions as harmful. And when we say that someone has a "good personality" or a "difficult personality," we mean that the person's consciousness, thought, rationality, and agency are manifested in pleasing or displeasing or bewildering ways that are particular to that individual. We do not ordinarily assess a being who lacks any one of these components of personhood in terms of their personality at all. Persons, then, express their innate personhood in their empirical personalities.

With these stipulations in place, I now turn to an analysis of the concept of xenophobia. Xenophobia is not simply an indiscriminate fear of strangers in general: it does not include, for example, fear of relatives or neighbors whom one happens not to have met. It is more specific than that. Xenophobia is a fear of individuals who violate one's empirical conception of persons and so one's self-conception. So xenophobia is an alarm reaction to a threat to the rational coherence of the self, a threat in the form of a theoretically anomalous other who transgresses our preconceptions about people. It is a paradigm example of reacting self-protectively to anomalous data that violate a provincial moral theory (an empirical conception of persons is inherently moral), by excluding the anomalous other from the realm of moral concern. In essence, the xenophobe violates the criteria of inclusiveness by failing to recognize and treat the anomalous other as fully a person.

Recall that on the proposed Kantian conception of the self elaborated in Chapter II, if we cannot make sense of such data in terms of those familiar concepts, we cannot register it as an experience at all. Recall also the argument of Chapter VII, that pseudorationality is an attempt to make sense of such data under duress, i.e. to preserve the internal rational coherence of the self, when we are baldly confronted by anomaly but are not yet prepared to revise or jettison our conceptual scheme accordingly. We saw that it is in the attempt to make sense of anomalous data in terms of empirically inadequate concepts that the mechanisms of pseudorationality – rationalization, dissociation and denial – kick in to secure literal self-preservation.

A familiar xenophobic example of *rationalization* would be conceiving of a slave imported from Africa as three-fifths of a person. This results from magnifying the properties that appear to support this diminished concept of personhood – the slave's environmental and psychological disorientation, lack of mastery of a foreign language, lack of familiarity with local social customs, incompetence at unfamiliar tasks, etc.; and minimizing the properties that disconfirm it – her capacity to learn, to forge innovative modes of communication and expression, to adapt and flourish in an alien and lethal social environment, to survive enslavement and transcend violations of her person, etc.

A xenophobic example of *dissociation* would include identifying Jews as subhuman, blacks as childlike, women as irrational, gays as perverts, or working class people as animals. This conceives of them as lacking essential properties of personhood, and so are ways of defining these groups of individuals out of our empirical conceptions of people. Similarly, xenophobic *denial* might include ignoring a woman's verbal contributions to a discussion, or passing over an African American's intellectual achievements, or forgetting to make provisions at a Christmas celebration for someone who is a practicing Jew. These are all ways of eradicating from one's domain of awareness properties that distinguish others as different from oneself.

Thus through the pseudorational mechanisms of rationalization and dissociation, xenophobia engenders various forms of stereotyping – racism, misogyny, anti-Semitism, homophobia, class elitism – that are discriminatory in both the cognitive and the political sense. It selects certain perceptually familiar properties of the person for disparagement, and distorts or

obliterates those that remain. It thereby reduces the complex singularity of the other's properties to an oversimplified but conceptually manageable subset, and this in turn diminishes one's full conception of personhood. This results in a provincial self-conception and conception of the world, from which significant available data are excluded. Thus violating inclusiveness reduces the scope of application of one's moral theory to the severely provincial. For its terms pick out only a small subset of the agents to whom in fact the theory applies. This provincial theory is then sustained with the aid of denial, by enforcing those stereotypes through such tactics as exclusion, ostracism, scapegoating, tribalism, and segregation in housing, education or employment. My thesis is that xenophobia is the originating phenomenon to which each of these forms of political discrimination is a response. The need for criteria of inclusiveness for a practically adequate moral theory arises from the operations of these pseudorational mechanisms in the social context.

Nevertheless, even if it is true that we are innately cognitively disposed to respond to any conceptual and experiential anomaly in this way, it does not follow that our necessarily limited empirical conception of people must be so limited and provincial as to invite it. A person could be so cosmopolitan and intimately familiar with the full range of human variety that only The Alien would rattle him. On the other hand, his empirical conception of people might be so limited that any variation in race, nationality, gender, sexual preference, or class would be cause for panic. How easily one's empirical conception of people is violated is one index of the scope of one's xenophobia; how central and pervasive it is in one's personality is another. In what follows I focus primarily on cases of political discrimination midpoint between such extremes: for example, of a European American who is thoughtful, well-rounded and well-read about the problems of racism in the United States, but who nevertheless feels fearful at being alone in the house with an African American television repairman. In all such cases, the range of individuals in fact identifiable as persons is larger than the range of individuals to whom one's empirical conception of people apply. In all such cases political discrimination can be understood in terms of certain corrigible cognitive errors that characterize prereflective xenophobia.

3. Failures of Cognitive Discrimination

By *cognitive discrimination*, I mean what we ordinarily understand by the term "discrimination" in cognitive contexts: A manifest capacity to distinguish veridically between one property and another, and to respond appropriately to each. When we say of someone that she is a discriminating person, for example, or that she has discriminating judgment, we mean, in part, that she is a person of refined tastes or subtle convictions; that she exercises a capacity to make fine distinctions between properties of a thing, and bases her positive or negative valuations on these actual properties. By contrast, when we say of a person that he lacks discrimination, we mean that he is unable to discern subtle differences or make fine distinctions; that he conflates properties or states of affairs that deserve separate consideration; or confuses qualities or ideas that are different. Thus lack of discrimination is often associated with aesthetic inadequacy; with an inability to discern quality – or, conversely, to appreciate the lowbrow on its

own terms. The following analysis mostly abstracts from these aesthetic associations – without, however, disavowing their relevance to an assessment of moral character. I take up the aesthetic dimension of failures of cognitive discrimination in Section 7, below.

3.1. The Error of Confusing People with Personhood

Xenophobia is fueled by a perfectly general condition of subjective consciousness, namely the first-/third-person asymmetry: Although I must identify myself as a person because of my necessary, enduring first-personal experience of rationally unified selfhood, my experience of you as a person, necessarily lacking that first-personal experience, can have no such necessity about it. Kant says it best:

> Identity of person is ... in my own consciousness unfailingly to be found. But when I view myself from the standpoint of another (as object of his outer intuition), this external observer considers me first and foremost in time .... So from the I, which accompanies all representations at all times in my consciousness, and indeed with full identity, whether he immediately concedes it, he will not yet conclude the objective continuity of my self. For because the time in which the observer situates me is not the same as that time to be found in my own, but rather in his sensibility, similarly the identity that is necessarily bound up with my consciousness, is not therefore bound up with his, i.e. with the outer intuition of my subject. (1C, A 362-363)

Kant is saying that the temporal continuity I invariably perceive in my own consciousness is not matched by any corresponding temporal continuity I might be supposed to have as the object of someone else's consciousness. Since I am not always present to another as I am to myself, I may appear discontinuously to her consciousness in a way I cannot to my own. And similarly, another may appear discontinuously to my consciousness in a way I cannot to my own.

Thus although personhood is a necessary concept of mine, whether or not any other empirical individual instantiates it is itself, from my point of view, a contingent matter of fact – as is the concept of that particular individual herself. Though you may exhibit rationality in your behavior, I may not know that, or fail to perceive it, or fail to understand it. Nor can you be a necessary feature of my experience, since I might I ignore or overlook you, or simply fail to have any contact with you. In any of these cases, you will fail to instantiate my concept of personhood in a way I never can. Because the pattern of your behavior is not a necessary and permanent, familiar concomitant of my subjectivity in the way my own unified consciousness and intellective processes are, I may escape your personhood in a way that I cannot escape my own. For me the innate idea of personhood is a concept that applies necessarily to me, but from my perspective, only contingently and empirically to you. Hence just as our experience of the natural world is limited relative to the all-inclusive, transcendent idea of its independent unity, similarly our empirical experience of other persons is limited relative to our all-inclusive, transcendent idea of personhood.

But there is an important disanalogy between them that turns on the problem of other minds and the first-/third-person asymmetry. For any empirical experience of the natural world we have, we must, according to Kant, be able to subsume it under the transcendent concept of a unified system of nature of which it is a part, even if we do not know what that system might be. By contrast, it is not necessarily the case that for any empirical experience of other people we have, we must be able to subsume them under the transcendent idea of personhood. This is because although they may, in fact, manifest their personhood in their personality, we may not be able fully to discern their personhood through its empirical manifestations, if those manifestations fall outside our empirical conception of what people are like.

Suppose, for example, that within my subculture, speech is used to seek confirmation and promote bonding, whereas in yours it is used to protect independence and win status;[2] and that our only interpersonal contact occurs when you come to fix my TV. I attempt to engage you in conversation about what is wrong with my TV, to which you react with a lengthy lecture. To you I appear dependent and mechanically incompetent, while to me you appear logorrheic and socially inappropriate. Each of us perceives the other as deficient in some characteristic of rationality: you perceive me as lacking in autonomy and basic mechanical skills, whereas I perceive you as lacking in verbal control and basic social skills. To the extent that this perceived deficit is not corrected by further contact and fuller information, each of us will perceive the other as less of a full-fledged person because of it. This is the kind of perception that contributes to one-dimensional stereotypes, for example of women as flighty and incompetent or of men as aggressive and barbaric, which poison the expectations and behavior of each toward the other accordingly. This is one way in which gender becomes a primary disvalued property.

By a *primary disvalued property* I mean a particular property of a person, seen as a source of intrinsic disvalue or incompetence, that is in fact irrelevant to judgments of that person's intrinsic value or competence – for example her race, gender, class, sexual orientation, or religious or ethnic affiliation. Conversely, I call any such arbitrary property perceived as a source of value or superiority a *primary valued property*.

Or take another example, in which the verbal convention in my subculture is to disclose pain and offer solace, whereas in yours it is to suppress pain and advert to impersonal topics; and that our only interpersonal contact occurs when I come to work as your housemaid.[3] Again each of us perceives the other as deficient in some characteristic of rationality: you perceive me as dull and phlegmatic in my lack of responsiveness to the impersonal topics you raise for discussion, whereas I perceive you as almost schizophrenically dissociated from the painful realities that confront us. Again, unless this perceived deficit is corrected by further contact and fuller information, each of us will perceive the other as less of a person because of it, thereby contributing to one-dimensional stereotypes of, for example, African Americans as stupid and of European Americans as ignorant and out of touch with reality; or of men as alienated and women as "in touch with their feelings," that similarly poison both the expectations and the

behavior of each toward the other. These are some further ways in which race or gender become primary disvalued properties.

In such cases there are multiple sources of empirical error. The first one is our respective failures to discriminate cognitively between the possession of rationality as an active capacity in general, and particular empirical uses or instantiations of it under a given set of circumstances and for a given set of ends. Because your particular behavior and ends strike me as irrational, I surmise that you must be irrational. Here the error consists in equating the particular set of empirical behaviors and ends with which I am familiar from my own and similar cases with unified rational agency in general. It is as though I assume that the only rational agents there are are the particular people I identify as such. Kant might put the point by saying that each of us has conflated his empirically limited conception of people with the transcendent concept of personhood.

3.2. The Error of Assuming Privileged Access to the Self

But now suppose we each recognize at least the intentionality of the other's behavior, if not its rationality. Since each of us equates rational agency in general exclusively with the motives and actions of her own subculture in particular, each also believes that the motives and ends that guide the other's actions – and therefore the evidence of conformity to the rule and order of rationality – nevertheless remain inaccessible in a way we each believe our own motives and ends not to be inaccessible to ourselves. This third-personal opacity yields the distinction between the appearance and the reality of the self: You, it seems, are an appearance to me behind which is hidden the reality of your motives and intentions, whereas I am not similarly an appearance that hides my own from myself. The less familiar you are to me, the more hidden your motives and intentions will seem, and the less benevolent I will assume them to be.

Of course whom we happen to recognize as familiar determines whose motives are cause for suspicion and whose are not. There is no necessary connection between actual differences in physical or psychological properties between oneself and another, and the epistemic inscrutability we ascribe to someone we regard as theoretically anomalous. It is required only that the other *seem* anomalous relative to our familiar subculture, however cosmopolitan that may be, in order to generate doubts and questions about what it is that makes him tick. Stereotypes of women as enigmatic or of Asians as inscrutable or of African Americans as evasive all express the underlying fear of the impenetrability of the other's motives. And someone who conceives of Jews as crafty, Africans as lazy, or women as devious expresses particularly clearly the suspicion and fear of various third-personal anomalous others as mendacious manipulators that is consequent on falsely regarding them as more epistemically inaccessible to one than one is to oneself.

Thus our mutual failure to identify the other as a person of the same status as oneself is compounded by scepticism based on the belief that each of us has the privileged access to his own personhood that demonstrates directly and first-personally what personhood really is. The

inaccessibility and unfamiliarity of the other's conception of her own motives to our consciousness of her may seem conclusive justification for our reflexive fear and suspicion as to whether her motives can be trusted at all.

Now Kant argues (1C, B 68-69, 153-156, 157-158 fn., A 551/B 579 fn.; G, Ak. 407) that from the first-personal relation I bear to my empirical self-conception which I lack to yours, it does not follow that my actual motives are any more accessible to me than yours are. Therefore, regardless of how comfortable and familiar my own motives may seem to me, it does not follow that I can know that my own motives are innocuous whereas yours are not. In fact, as Nietzsche argues, it is difficult to imagine how I might gain any understanding of the malevolent motives I reflexively ascribe to you at all, without having first experienced them in myself. Of course this is not to say that I cannot understand what it means to be the victim of maleficent *events* without having caused them myself. But it is to say that I must derive my understanding of the malevolent *intentionality* I ascribe to you from my own first-hand experience of it. Therefore your epistemic opacity to me furnishes no evidence for my reflexive ascription to you of malevolent or untrustworthy motives, although that ascription itself does furnish evidence for a similar ascription of them to myself. Thus Kant might put this second error by saying that we have been fooled by the first-/third-person asymmetry into treating the ever-present "dear self" as a source of genuine self-knowledge on the basis of which we make even faultier and more damaging assumptions about the other.

3.3. The Error of Failing to Modally Imagine Interiority

These two errors are interconnected with a third, namely our respective failures to modally imagine each other's behavior as animated by the same interior elements of personhood that animate our own, i.e. consciousness, thought, and rationality. Our prior failure to recognize the other's behavior as manifesting evidence of interiority – a failure compounded by conceptual confusion and misascription of motives – then further undermines our ability to bridge the first-/third-person asymmetry by imagining the other to have them. Since, from each of our first-personal perspectives, familiar empirical evidence for the presence of interiority is lacking in the other, we have no basis on which to make the ascription, and so no basis for modally imagining what it must be like from the other's perspective. Our respective, limited empirical conceptions of people, then, itself the consequence of ignorance of others who are thereby viewed as different, delimits our capacity for empathy. This is part of what is involved in the phenomenon feminists refer to as objectification, and what sometimes leads men to describe women as self-absorbed. Kant might put this point by saying that by failing to detect in the other's behavior the rule and order of rationality that guides it, we fail to surmise or imagine the other's motives and intentions.

This error, of failing to modally imagine the other as similarly animated by the psychological dispositions of personhood, is not without deleterious consequences for the xenophobe himself. In Chapter VI.2 I described the egocentric and narrowly concrete view of the

world that results from the failure to imagine empathically another's interiority, and its interpersonal consequences. From the first-personal perspective, this error compounds the seeming depopulation of the social environment of persons and its repopulation by impenetrable and irrational aliens. This is to conceive one's social world as inhabited by enigmatic and unpredictable disruptions to its stability, to conjure chimaeras of perpetual unease and anxiety into social existence. Relative to such a conception, segregation is no more effective in banishing the threat than is leaving on the nightlight to banish ghosts, since both threats arise from the same source. Vigilance and a readiness to defend oneself against the hostile unknown may become such intimately familiar and constitutive habits of personality that even they may come to seem necessary prerequisites of personhood.

4. Test Case #4: Political Discrimination

The three foregoing errors involve failures of cognitive discrimination for which a well-intentioned individual could correct. For example, someone who regularly confuses people with personhood might simply take a moment to formulate a general principle of rational behavior that both applies to all the instances with which she is familiar from her particular community and has broader application as well; and remind herself, when confronted by theoretically anomalous behavior, to at least try to detect the operation of that principle within it. Similarly, it does not require excessive humility on the part of a person who falsely assumes privileged access to the self to remind himself that our beliefs about our own motives, feelings, and actions are exceedingly fallible and regularly disconfirmed; and that it is therefore even more presumptuous to suppose any authority about someone else's. Nor is it psychologically impossible to gather information about others' interiority – through research, appreciation of the arts, or direct questioning and careful listening, so as to cultivate one's imaginative and empathic capacities to envision other minds. Thus it is possible for someone to have such xenophobic reactions without being a full-blown xenophobe, in the event that she views them as causes for concern rather than celebration. She may experience these cognitive failures without being a first-order political discriminator, in the event that she has no personal investment in the defective empirical conception of people that results; and is identifiable as a bona fide first-order political discriminator to the extent that she does.

By *political discrimination*, I mean what we ordinarily understand by the term "discrimination" in political contexts: A manifest attitude in which a particular property of a person that is irrelevant to judgments of that person's intrinsic value or competence, for example his race, gender, class, sexual orientation, or religious or ethnic affiliation, is seen as a source of intrinsic disvalue or incompetence; in general, as a source of inferiority.

Just as, for reasons of simplicity and structure I restricted my analysis of an ideal descriptive moral theory to Theory *K* in Chapters V.5.1 – 2 and IX, here I restrict my analysis of political discrimination to consideration of *intrinsic* value or competence, for similar reasons. Thus I ignore considerations of instrumental value or competence in furthering some specified

social or institutional policy, of the sort that would figure in arguments that would justify, e.g., hiring someone as a role model in a classroom, or to provide a unique and needed perspective in a business venture or court of law; or, on the other hand, hiring someone to a professional position solely in order to meet affirmative action quotas; or refusing to sell real estate in a certain neighborhood to an African American family because doing so would lower property values; or refusing to serve Asians at one's family diner because it would be bad for business.

I distinguish between two kinds of political discrimination: *first-order political discrimination* as defined above, and *higher-order political discrimination* as a refinement introduced by pangs of conscience that result in even more radical failures of cognitive discrimination: of the other, of oneself, and of the situation. Judging a person as inferior because one perceives her race as a primary disvalued property depends upon failing to distinguish finely enough between properties she has and those she does not have, and between those which are relevant to such a judgment and those which are not. This is the essence of xenophobia. Our inability to make fine-grained cognitive discriminations in judging a person is the result of a fear reaction to the theoretically anomalous perceptual data that person presents, and the cause of a corresponding inability to evaluate her non-pseudorationally as a person.

4.1. First-Order Political Discrimination

A person could make the first three cognitive errors described in Sections 3.1 – 3.3 above without taking any satisfaction in his provincial conception of people ("Is this really all there is?" he might think to himself about the inhabitants of his small town), without identifying with it (he might find them boring and feel ashamed to have to count himself among them), and without feeling the slightest reluctance to enlarge and revise it through travel or exploration or research.

What distinguishes a first-order political discriminator is her personal investment in her provincial conception of people. Her sense of literal self-preservation requires her conception to be viridical, and is threatened when it is disconfirmed. She exults in the thought that only the people she knows and is familiar with (whites, blacks, WASPs, Jews, residents of Crawford, Texas, members of the club, etc.) are persons in the full, honorific sense. This is the thought that motivates the imposition of politically discriminatory stereotypes, both on those who confirm it and those who do not.

To impose a *stereotype* on someone is to view him as embodying a limited set of properties falsely taken to be exclusive, definitive, and paradigmatic of a certain kind of individual. I shall say that a stereotype

> (a) equates one contingent and limited set of primary valued properties that may characterize persons under certain circumstances with the universal concept of personhood;
>
> (b) restricts that set to exclude divergent properties of personhood from it;
>
> (c) withholds from those who violate its restrictions the essential properties of personhood; and

(d) ascribes to them the primary disvalued properties of deviance from it.

Thus a stereotype identifies as persons those and only those who manifest the primary valued properties in the set ((a) and (b)), and subsidiary ones consistent with it (such as minor personality quirks or mildly idiosyncratic personal tastes). Call this set the *honorific stereotype*, and an individual who bears such primary valued properties the *valuee*. And reciprocally, the honorific stereotype by implication identifies as deviant or anomalous all those who manifest any properties regarded as inconsistent with it ((c) and (d)). Call this second set of primary disvalued properties the *derogatory stereotype*, and an individual who bears such primary disvalued properties the *disvaluee*.

So, for example, an individual who bears all the primary valued properties of the honorific stereotype as required by (a) may be nevertheless disqualified for status as a valuee according to (b), by bearing additional primary disvalued ones as well. She may be related by blood or marriage to a Jew, for example; or have bisexual inclinations; or, in the case of an African American, an enthusiasm for classical scholarship. In virtue of violating (b), one may then fail to qualify as a full-fledged person at all (c), and therefore may be designated as deviant by the derogatory stereotype according to (d). The derogatory stereotype most broadly includes all the primary disvalued properties that fall outside the set defining the honorific stereotype (i.e. "us versus them"), or may sort those properties into more specific subsets according to the range of individuals available for sorting.

A politically discriminatory stereotype generally is therefore distinguishable from an inductive generalization by its provincialism, its oversimplification, and its rigid imperviousness to the complicating details of singularity. Perhaps most importantly, a discriminatory stereotype is distinguishable from an inductive generalization by its function. The function of an inductive generalization is to guide further research, and this requires epistemic alertness and perceptual sensitivity to the possibility of confirming or disconfirming evidence in order to make use of it. An inductive generalization is no less a generalization for that: it would not, for example, require working class African Americans living in the Deep South during the 1960s to dismantle the functionally accurate and protective generalization that white people are dangerous. What would make this an inductive generalization rather than a stereotype is that it would not preclude recognition of a European American who is safe should one appear. By contrast, the function of a stereotype is to render further research unnecessary. If the generalization that white people are dangerous were a stereotype, adopting it would make it cognitively impossible to detect any European Americans who were not.

Thus Kant might describe the reciprocal imposition of stereotypes as the fallacy of equating a partial and conditional series of empirical appearances of persons with the absolute and unconditioned idea of personhood that conceptually unifies them. Whereas cognitive failure 3.1 – of confusing one's empirical conception of people with the transcendent concept of personhood – involves thinking that the only persons there are are the people one knows, this fourth error – of equating personality with personhood – involves thinking that the kind of

persons one knows are all there can ever be. So unlike inductive generalizations, the taxonomic categories of a stereotype are closed sets that fundamentally require the binary operation of sorting individuals and properties into those who fall within them and those who do not.[4]

As a consequence of his personal investment in an honorific stereotypical conception of persons, a first-order political discriminator has a personal investment in an honorific stereotypical self-conception that is therefore self-aggrandizing in the sense explained in Chapter VII. This means, to review, that this self-conception is a source of personal satisfaction or security to him; that to revise or disconfirm it would elicit in him feelings of dejection, deprivation or anxiety; and that these feelings are to be explained by his identification with this self-conception. In order to maintain his honorific and self-aggrandizing self-conception, a first-order political discriminator must perform the taxonomic binary sorting operation not only on particular groups of ethnic or gendered others, but on everyone, including himself. Since his self-conception as a person requires him and other *bona fide* persons to dress, talk, look, act, and think in certain highly specific and regimented ways in order to qualify for the honorific stereotype, everyone is subject to scrutiny in terms of it.

This is not only prejudicial to a disvaluee who violates these requirements and thereby earns the label of the derogatory stereotype. It is also prejudicial to a valuee who satisfies them, just in case there is more to his personality than the honorific stereotype encompasses and more than it permits. Avoidance of the negative social consequences of violating the honorific stereotype – ostracism, condemnation, punishment, or obliteration – necessitates stunting or flattening his personality in order to conform to it (for example, by eschewing football or nightclubs, and learning instead to enjoy scholarly lectures as a form of entertainment because one is given to understand that that is the sort of thing real intellectuals typically do for fun); or bifurcating his personality into that part which can survive social scrutiny and that "deviant" part which cannot (as, for example, certain government officials have done who deplore and condemn homosexuality publicly on the one hand, while engaging in it privately on the other). One reason it is important not to equate personality with personhood is so that the former properties can flourish without fear that the latter title will be revoked.

Truncating his personality in order to conform to an honorific stereotype in turn damages the political discriminator's self-esteem and also his capacity for self-knowledge. Someone who is deeply personally invested in the honorific stereotype but fails fully to conform to it (as everyone must, of course) views himself as inherently defective. He is naturally beset by feelings of failure, inferiority, shame and worthlessness which poison his relations with others in familiar ways: competitiveness, dishonesty, defensiveness, envy, furtiveness, insecurity, hostility, and self-aggrandizement are just a few of the vices that figure prominently in his interpersonal interactions. But if these feelings and traits are equally antithetical to his honorific stereotype, then they, too, threaten his honorific stereotypical self-conception and so are susceptible to pseudorational denial, dissociation or rationalization. For example, a first-order political discriminator might be blindly unaware of how blatantly he advertises these feelings and traits in

his behavior; or he might dissociate them as mere peccadilloes, unimportant eccentricities that detract nothing from the top-drawer person he essentially is. Or he might acknowledge them but rationalize them as natural expressions of a Nietzschean, *übermenschliche* ethic justified by his superior place in life. Such pseudorational habits of thought reinforce even more strongly his personal investment in the honorific stereotype that necessitated them, and in the xenophobic conception of others that complements it. This fuels a vicious downward spiral of self-hatred and hatred of anomalous others from which it is difficult for the political discriminator to escape. Thus the personal disadvantage of first-order political discrimination is not just that the discriminator devolves into an uninteresting and malevolent person. He damages himself for the sake of his honorific stereotype, and stunts his capacity for insight and personal growth as well.

A sign that a person's self-aggrandizing self-conception is formed by an honorific stereotype is that revelation of the deviant, primary disvalued properties provokes shame and denial, rather than a reformulation of that self-conception in such a way as to accommodate them. For example, a family that honorifically conceives itself as white Anglo-Saxon Protestant may deny that its most recent offspring in fact has woolly hair or a broad nose. Similarly, a sign that a person's conception of another is formed by a derogatory stereotype is that revelation of the other's nondeviant, primary valued properties provokes hostility and denial, rather than the corresponding revision of that conception of the other in such a way as to accommodate them. For example, a community of men that honorifically conceives itself in terms of its intellectual ability may dismiss each manifestation of a woman's comparable intellectual ability as a fluke.

These two reactions are reciprocal expressions of the same dispositions in the first- and third-person cases respectively. Shame involves the pain of feeling publicly exposed as defective, and denial is the psychological antidote to such exposure: for example, if the purportedly WASP offspring does not have negroid features, there is nothing for the family to feel ashamed of. So a person whose self-aggrandizing self-conception is defined by an honorific stereotype will feel shame at having primary disvalued properties that deviate from it, and will attempt to deny their existence to herself and to others. By contrast, hostility toward another's excellence is caused by shame at one's own defectiveness, and denial of the excellence is the social antidote to such shame: for example, if the woman is not as intelligent as the men are purported to be, then there is no cause for feeling shamed by her, and so none for hostility toward her. So a person whose self-aggrandizing self-conception is formed by an honorific stereotype will feel hostility toward a disvaluee who manifests valued properties that violate the derogatory stereotype imposed on him; and will attempt to deny the existence of those valued properties in the other to herself and to others.

In the first-person case, the objects of shame are primary disvalued properties that deviate from one's honorific stereotypical self-conception. In the third-person case, the objects of hostility are valued properties that deviate from one's derogatory stereotypical conception of the disvaluee. But in both cases the point of the reactions is the same: to defend one's stereotypical self-conception against attack, both by first-person deviations from it and by third-person

deviations from the reciprocal stereotypes this requires imposing on others. And in both cases, the xenophobic reactions are motivated in the same way: the properties regarded as anomalous relative to the stereotype in question are experienced by the first-order political discriminator as an assault on the rational coherence of his theory of the world – and so, according to Kant, on the rational coherence of his self.

Indeed, left untreated, all four of these cognitive failures more generally – the conflation of the transcendent concept of personhood with one's provincial conception of people that another happens to violate, the ascription to the other of malevolent motives on the basis of an epistemically unreliable self-conception, the inability to imagine the other as animated by familiar or recognizably rational motives, and the equation of personality with personhood inherent in the imposition of reciprocal stereotypes – combine to form a conception of the other as an inscrutable and malevolent moral anomaly that threatens that provincial moral theory which unifies one's experience and structures one's expectations about oneself and other people. If this were an accurate representation of others who are different, it would be no wonder that xenophobes feared them.

4.2. Reciprocal First-Order Political Discrimination

So far I have argued that first-order political discrimination involves the reciprocal imposition of honorific and derogatory stereotypes, on oneself and on the theoretically anomalous other respectively. But is it not possible to value properties ordinarily taken to be irrelevant to judgments of a person's value or competence without eliciting the charge of honorific stereotyping? Are such primary valued properties ever relevant to judgments of a person's noninstrumental value or competence? By *reciprocal first-order political discrimination*, I mean a manifest attitude in which a particular property of a person that is irrelevant to judgments of that person's noninstrumental value or competence, for example her race, gender, class, sexual orientation, or religious or ethnic affiliation, is seen as a source of value or competence, in general, as a source of superiority. Primary valued properties are those perceived as elevating and valorizing its bearers accordingly.

Take the case in which we are particularly drawn to befriend a valuee with whom we share a similar ethnic background, because we expect to have more in common (lifestyle, tastes, sense of humor), share similar values, or see the world from a similar perspective. In this kind of case the primary valued property is not, say, being Jewish; but rather having the same ethnic background, whatever that may be. Is similarity of ethnic background a property that is relevant to our judgments of how valuable the valuee is as a friend? No, for it does not form any part of the basis for such a judgment. That a friendship is better, richer, or more valuable in proportion to the degree of similarity of the friends' ethnic backgrounds is a judgment few would be tempted to make.

In these cases, it is not the valuee's similar ethnicity itself that is the source of value, but rather the genuinely valuable properties – for example, similarity of values or worldview – with

which we expect similar ethnicity to be conjoined. Rather than making a normative judgment about his value or competence as a friend in this case, we in fact make an epistemic judgment about the probability that, given the valuee's ethnic identity, he will bear properties susceptible of such normative judgments. These epistemic rules of thumb are defeasible, and may have disappointing consequences for personal relationships. For they ascribe primary value to a kind of property at the expense of others that are in fact more important for friendship – like sensitivity, similarity of values, tastes or experiences, or mutual respect – with which that kind of property is only contingently, if ever, conjoined. Presumably something like this may explain the malaise of someone who has chosen all the "right" friends, married the "right" spouse, and landed the "best" job, yet feels persistently unhappy, disconnected, and dissatisfied in her social relationships.

If similarity of race, gender, sexual orientation, class background, or religious or ethnic affiliation are in themselves irrelevant to judgments of a person's noninstrumental value or competence, primary valued properties such as being of a particular race, gender, etc. are even more obviously so. At least it has yet to be demonstrated that any particular racial, ethnic, gender, class or religious group possesses the properties necessary for, e.g., friendship to an outstanding degree. The thesis that women make better friends is often supported by arguments to the effect that they become closer confidantes more quickly. But there are many other properties that contribute to friendship – trustworthiness, loyalty, dependability, honesty, mutual respect, etc. – that such arguments ignore. Epistemic probability judgments about the concatenation of any such primary valued properties with genuinely valuable traits, such as sensitivity or similarity of interests, also may bias our ability to perceive clearly the properties a particular individual actually has – as when a wife minimizes the reality and seriousness of her husband's physical abuse of her, because of the weight she accords his class background. This would be a case of reciprocal first-order discrimination, according to the above definition, because she sees a primary valued property – class background – that is irrelevant to judgments of the valuee's noninstrumental value or competence as a spouse as a (compensating) source of superiority.

It might be objected that such epistemic rules of thumb are inductive generalizations, however irrational or poorly grounded, that we need in order to survive in a world of morally opaque others: How ought one behave, for example, alone in a subway car with four hooded African American male teenagers carrying ghetto blasters and wearing running shoes? However, even if it were true that most muggers were hooded African American male teenagers in running shoes, it still would not follow that most hooded African American male teenagers in running shoes were muggers. This epistemic rule of thumb is a stereotype, not an inductive generalization, if it leads one to react to every hooded African American male teenager in running shoes one encounters as though he were a mugger when there is no independent justification for thinking he is.

Alternately, one may make a judgment of value about some such property abstractly and independently considered. One may value being African American, or Irish American, or of working class origins, for its own sake. Or one may choose a partner from the same religion because one views that religion and its traditions themselves as intrinsically valuable, independently of one's partner's compatibility with respect to lifestyle, values, or worldview. Here the judgment of value is directed not at the valuee's value or competence, but rather at the property she bears and to the preservation or affirmation of which one's choice of her is instrumental. Nothing in the following discussion addresses or precludes such judgments, although there is much to say about them. My target is judgments of noninstrumental value about individuals, not about properties of individuals abstractly and independently considered, to which individuals themselves are instrumental.

Is it humanly possible to value a person just and only because he bears some such primary valued property – not because of the further properties with which we expect that one to be conjoined, but just for the sake of that property in itself? It is difficult to make sense of this. Suppose I value Germanness because the Germans I have known tend to have deep passions and an amusingly fatalistic sense of humor; and that I then meet a shallow and phlegmatic German with no sense of humor at all. In the absence of other, unexpectedly attractive personality characteristics I may appreciate, just what is it about being German in itself that is supposed to confer worth on this particular individual? Either we must be able to spell out an answer to this question in terms of other properties that are only contingently connected, if at all, to this one – for example, having been socialized within a certain culture "from the inside", being part of a certain historical tradition, etc. – or else we are appealing to a mysterious and ineffable, non-natural property of Germanness.

For purposes of this discussion I ignore the range of cases in which my valuation of, for example, Germanness is rooted in the status or worth I expect my choice of German friends to confer on *me*. This may occur either where the primary valued property is one shared by oneself, or where it is not. Thus it may happen that one's choice of a European American, Anglo-Saxon Protestant spouse is made in part with an eye to reinforcing to others and to oneself the primary value of one's own status as a European American, Anglo-Saxon Protestant. Or, alternately, one's contrasting choice of an African American, Methodist spouse may be made with an eye to proving to others and to oneself one's "cool", sophistication, or commitment to civil rights. These are all cases in which the property is valued as a source of instrumental value or competence, namely for its ability to confer value on the reciprocal first-order political discriminator. So I leave them aside here.

Then suppose there are ineffable, non-natural properties such as Germanness, and that we may arguably appeal to them. To what degree might Germanness outweigh the person's other properties that, by hypothesis, I deplore? Surely the mere fact of Germanness can provide no consolation at all, in practice, for other properties of the person that offend me. It will not compensate, for example, for a failure to laugh at my jokes, or a tendency to discuss the weather

at excessive length, or to fall asleep at the opera. And then it is hard to see in what its purported value consists.

Independently of the other, genuinely valuable properties with which they are only contingently, if at all, conjoined, properties such as race, gender, sexual orientation, class background, or religious or ethnic affiliation are in themselves always irrelevant to judgments of a person's noninstrumental value or competence. This holds whether they are considered as primary disvalued or valued properties, and even where they are used as epistemic rules of thumb for detecting such properties. We may in fact feel compelled to make such judgments, in the service of expediency, or what we imagine to be our self-interest, and screen our circle of associates accordingly. But it is nothing to be proud of.

4.3. Higher-Order Political Discrimination

Next I examine a more sophisticated manifestion of political discrimination that is supervenient on the first-order political discrimination just discussed. I shall call this *higher-order political discrimination*. As in first-order political discrimination, a higher-order discriminator manifests in behavior the attitude in which a particular property of a person that is irrelevant to judgments of that person's intrinsic value or competence, e.g. her race, gender, class, sexual orientation, or religious or ethnic affiliation, is seen as a source of disvalue or inferiority, i.e. as a primary disvalued property. By *second-order political discrimination*, I will understand the attitude within which a primary disvalued or valued property in turn confers disvalue or value respectively on further properties of the disvaluee or valuee respectively. I shall refer to these latter as *secondary disvalued (or valued) properties*.

4.3.1. Transitivity and Comprehensiveness

Second-order political discrimination works in the following way. A disvaluee's primary disvalued property, say, being a male homosexual, causes the second-order political discriminator to view some *further* property of the disvaluee, say, being an eloquent speaker, in a negative light. The respect in which this further property is seen as negative depends on the range of possible descriptions it might satisfy, as well as the context in which it appears. Thus, for example, the second-order political discriminator might view the disvaluee's eloquence as purple prose, or empty rhetoric, or as precious, flowery, or mannered. These predicates are not interchangeable for the second-order political discriminator. Nor are they taken to be arbitrarily applied. The second-order political discriminator will choose from among them to express his disvaluation in response to contingencies of the situation and individuals involved. He may, in all sincerity, explain his disvaluation with reference to impartially applied aesthetic standards, or to his ingrown, native suspicion of big words. But the crucial feature of second-order political discrimination is that the actual explanation for his disvaluing the person's eloquence, *in whatever respect he disvalues it*, is the person's primary disvalued property of being a male homosexual.

Does second-order political discrimination as thus defined ever actually occur? Some familiar examples of it include attaching disvalue to a person's having rhythm, by reason of its putative connection with her being black; or attaching disvalue to a person's being very smart, by reason of its putative connection with his being Jewish. Both of these cases are examples of politically discriminatory stereotyping, in which some arbitrary property is falsely taken to be characteristic of persons of a particular race or ethnic or religious affiliation. But I mean to call attention to a slightly different feature of these examples. Someone who practices second-order political discrimination regards a black person who has rhythm, as vulgar, salacious, or offensive; at the very least, undignified. Similarly, such a person regards a Jewish person who is very smart as sophistical, glib, or crafty, or as subversive or ungentlemanly; at the very least, untrustworthy. In both cases, properties that are in themselves salutary, or at least neutral, are castigated by the second-order political discriminator, by reason of the disvalue conferred on them by the primary disvalued property. This is what makes them examples of second-order political discrimination.

These familiar, stereotypic examples of second-order political discrimination do not exhaust the repertoire of higher-order political discrimination, for many reasons. First, orders of discrimination can, in theory, be multiplied indefinitely. So, for example, a case of *third-order political discrimination* would involve what I shall call *tertiary disvalued properties*: The primary disvalued property (say, being black) confers disvalue on a further, secondary disvalued property (having rhythm), which in turn confers disvalue on yet a further property of the person (say, being a good dancer). Having rhythm is seen as vulgar, by reason of its association with being black, and being a good dancer is then seen as exhibitionistic (say), by reason of its association with having rhythm. In any such case, the primary property is in fact irrelevant to judgments of a person's value or competence. Hence the value or disvalue it confers on secondary, tertiary, etc. properties is bogus.

The *n*-order disvalue relation is *transitive*, in that, for example, if being black confers disvalue on having rhythm, and having rhythm confers disvalue on being a good dancer, then being black confers disvalue on being a good dancer. The *n*-order disvalue relation is also *comprehensive*, in that the primary disvalued property poisons the higher-order political discriminator's evaluations of all further properties of the disvaluee. For example, the primary disvalued property of being black may confer disvalue, alternatively, on a dancer's classical styling: Classical styling in a black dancer may be seen as inappropriate, or as an obscene parody of traditional ballet. Of course there are other, more convoluted cases of higher-order political discrimination that represent epicyclic variations on the straightforward cases I examine here. For example, being black may wildly exaggerate the value attached to classical styling in a black dancer, if classical styling is perceived as something the person had to overcome great innate and cultural obstacles to achieve. In either case, being black functions as a primary disvalued property because it carries a presumption of inferiority into the evaluation of further properties of the person. The primary disvalued property also confers disvalue on other, unrelated properties of the disvaluee: her appearance, accent, mode of dress, etc.

Is it perhaps too strong to claim that a primary disvalued property poisons the higher-order political discriminator's evaluation of *all* of the disvaluee's other properties? Can't a higher-order political discriminator respect a disvaluee's traits of character in a certain restricted area, despite his disvalued status? I am inclined to think not. For this seems to occur almost exclusively when the "valued" property itself conforms to the higher-order political discriminator's stereotypes. For example, an African American man may be admired for his athletic prowess but encounter hostility when he runs for political office. In such cases, the higher-order political discriminator's admiration and respect for the stereotypical trait is not unalloyed. It is tempered by a certain smug complacency at the disvaluee's confirmation of his disvalued status in the very cultivation and expression of that stereotypical trait. To sustain the above objection, we would need to see a higher-order political discriminator exhibiting unalloyed admiration and respect for *non*stereotypical traits, such that these positive feelings did not, in turn, positively reform the higher-order political discriminator's prejudicial attitude toward the person's primary disvalued property: Someone who sincerely respects and admires a disvaluee for nonstereotypical reasons, without feeling threatened or invaded, has already begun to weaken the psychological edifice on which her politically discriminatory evaluation of the person as a disvaluee is based.

The comprehensiveness of the *n*-order disvalue relation underscores a second reason why stereotypical cases of second-order discrimination do not exhaust the repertoire of higher-order discrimination: Nonstereotypical traits are also recruited to receive value or disvalue from primary properties to suit particular occasions. We do not ordinarily think of classical styling in dance as a property about which discriminators might have any particular attitude. But this may be mistaken. Higher-order discrimination is not concerned solely with *stereotypical* secondary, tertiary, etc. disvalued properties. It may be concerned with *any* further properties of the person on which the primary disvalued property itself confers disvalue. Thus, for example, being Jewish (or Nigerian, or a woman) may confer disvalue on being smart, which in turn may confer disvalue on being intellectually prolific: A person's intellectual prolificity may be seen as evidence of logorrhea, or lack of critical conscience, and may thus poison the evaluation of those intellectual products themselves.

A first test for ascertaining whether the disvalue of some property of a person is to be explained as a case of higher-order political discrimination is to ascertain whether or not that property is disvalued uniformly across individuals, regardless of anything that might count as a primary disvalued property for a higher-order political discriminator. If someone is just as contemptuous of Fred Astaire's having rhythm as they are of Michael Jackson's, or just as contemptuous of intellectual prolificity in Balzac as in Isaac Asimov, then the charge of higher-order political discrimination may be defeated.

It might be thought that this first test is inherently self-limiting for the case in which the person happens to dislike just the property that is most typically associated with, e.g. a certain race – say, dark skin, but nevertheless passes the first test in that he disvalues it uniformly across

individuals, whether it occurs in Tanzanians, African Americans, Native Americans, Indians, Jews, Arabs, Aborigines, or tanning lotion-soaked Californians. This kind of case does not present a problem. The fact that someone is acquitted of being a racist does not imply that his evaluations are therefore admirable or enlightened. Any predicate or combination of predicates that fails the first test is either a rigged definite description of a particular disvalued group, for example, "ova-producing featherless bipeds," or else describes a discriminatory stereotype, e.g. "dark-skinned, dark-eyed, woolly-haired individuals with rhythm." Of course, a person might just happen to disvalue only individuals who fit such a stereotype and not those who violate it. But since this disvaluation would not be independent of anything that might count as a primary disvalued property for such a person, it would not defeat the charge of higher-order political discrimination.

Note, however, that the first test does not work for identifying a distinct but related attitude, which we might call *generalized higher-order political discrimination*, in which a person comes to disvalue some constellation of higher-order properties across the board specifically because of its original association with a primary disvalued property stereotypically ascribed to a certain group. Someone who finds having rhythm vulgar in any dancer, regardless of racial or ethnic affiliation, because she associates having rhythm with African Americans, whom she fears and despises, would exemplify such an attitude.

Stereotypes change in accordance with changes in the objects of political discrimination, as different populations seek access to the goods, services and opportunities enjoyed by the advantaged; and primary and higher-order disvalued properties change accordingly. For instance, the anti-Semitic response to the attempts of Jewish intellectuals to achieve full assimilation into the institutions of higher education in the United States frequently found expression in the disvaluative description of assertively ambitious Jewish academics as pushy or opportunistic. Now similarly situated African Americans, Asians and women frequently enjoy that title. Conversely, those with such primary disvalued properties who attempt to substitute diplomacy for assertion are characterized by higher-order political discriminators as manipulative, obsequious, or sycophantic.

A second test for ascertaining whether or not the disvalue of some property of a person is to be explained as a case of higher-order political discrimination is to ascertain whether there is any alternative property, conduct or manner, directed toward the same goal – i.e. of gaining access to unjustly withheld social advantages, that avoids or deflects the disvalue conferred by the primary disvalued property. If there is not – if, that is, whatever your strategy, you're damned if you do and damned if you don't, then the charge of higher-order political discrimination is *prima facie* justified.

Other arbitrary properties, not just the familiar political ones, can function as primary disvalued properties to a higher-order political discriminator. Physical appearance, style of diction, social bearing, familial, educational, or professional pedigree, circle of associates, manner of dress, are among the more familiar, if less widely acknowledged, objects of higher-order

political discrimination. Some of these properties are often assumed to go hand-in-hand with, or even be partially definitive of, more widely recognized primary disvalued properties. For example, higher-order political discriminators may tend to assume that ethnic identity is inherently connected with a certain physical appearance (Jews have dark, curly hair and long noses), that racial identity is connected with a certain style of diction and class background (African Americans speak Black English and come from the ghetto), or that gender identity is connected with a certain social bearing (women are sympathetic, passive, and emotional). This is how a stereotype is formed.

But again I mean to call attention to a slightly different point: These properties themselves may be seen as sources of disvalue, independently of their possible connection with such stereotypically primary disvalued properties. Someone who has all of the valued race, ethnic, religious, class, and gender properties, but lacks the valued style of diction, mode of self-presentation, or educational or professional pedigrees may be subject to higher-order political discrimination just as fully as someone who lacks all of the former properties but has all of the latter. In both cases, this means that their other properties – their personality characteristics, interests, or achievements – will be seen as higher-order disvalued properties, by reason of their association with these equally arbitrary primary disvalued properties.

This shows that the first-order political discrimination with which we are familiar is merely a special case of a more general psychological phenomenon that is not limited to first-order *political* discrimination at all. However, higher-order political discrimination as defined above usually includes it. For it would be psychologically unusual, to say the least, to find an individual who is in general corrupt in his evaluations of a person's other properties in the ways just described, yet impartial and scrupulous in his evaluations of blacks, Jews, women, gays, etc. and their properties. Someone who is apt to dislike a person because of her hair texture or accent or mode of dress can hardly be expected to be genuinely judicious when it comes to judging her gender, race, class, sexual orientation, or ethnic or religious affiliation. Hence we can expect that first-order political discrimination and higher-order political discrimination in general are to be found together.

There is another reason that favors retaining the label of higher-order political discrimination, despite its application to primary disvalued properties less widely recognized as political in nature, corresponding to a broader conception of political behavior. We can think of politically discriminatory stereotyping more generally as a means of sorting individuals into those with whom one is willing to share available power and resources versus those with whom one is not; that is, as a means of sorting others into those whom one accepts into one's moral community and those whom one excludes from it. In this broader sense, any disvalued property can become a criterion for excluding the disvaluee from the discriminator's circle of honorifically stereotyped valuees – rationality, pain, and insight among them.

4.3.2. Reciprocality

Higher-order political discrimination as so far described implies a companion phenomenon, which I shall call *reciprocal higher-order political discrimination*. Here properties irrelevant to judgments of a person's competence or worth are seen as primary *valued* properties, as sources of value that confer value on the person's secondary, tertiary, etc. properties. Any one of the primary properties enumerated so far may have this function. For example, a person's gender may be perceived as conferring value on secondary properties, such as his competence to hold a certain professional position. Or a person's familial lineage may be perceived as conferring value on her admissability to an institution of higher education. Or a person's class background may be perceived as conferring value on his manner of dress. Or a person's educational pedigree may be perceived as conferring value on her political pronouncements, which in turn confers value on her personal lifestyle; and so on. Each of these examples has an arbitrary and irrational quality to them. That is because reciprocal higher-order political discrimination, like higher-order political discrimination itself, is an arbitrary and irrational attitude.

Higher-order political discrimination and reciprocal higher-order political discrimination are materially interdependent. If a person's having a particular racial identity is a source of disvalue for a higher-order political discriminator, then if someone lacks that racial identity, he is not seen as tainted by that disvalue. For example, if a person's being Asian confers disvalue on his attempts at tact, i.e. if he is therefore perceived as particularly evasive and inscrutable, then if he were viewed as white, he would not be perceived as similarly evasive and inscrutable. For if a higher-order political discriminator recognized that one can be just as evasive and inscrutable without being Asian, say, if one has a hidden agenda or lacks social skills, then it would have to be recognized that those properties, rather than his being Asian, might be conferring disvalue on his attempts at tact. Conversely, if a person's having a particular racial identity is a source of value for a higher-order political discriminator, then someone who lacks that racial identity is not blessed by that value. For example, if a person's being viewed as white confers value on her attempts at tact, i.e. if she is therefore viewed as sensitive and reasonable, then if she were Asian, she would not be perceived as similarly sensitive and reasonable. For if a higher-order political discriminator recognized that one can be just as sensitive and reasonable without being white, say, if one has no personal investment in the issue or has thought hard about it, then it would have to be recognized that those properties, rather than her being white, might be conferring value on her attempts at tact.

The two tests for higher-order political discrimination apply analogously to reciprocal higher-order political discrimination: First, ascertain whether or not the higher-order valued property is valued uniformly across individuals, regardless of anything that might count as a primary valued property for the discriminator. If a person's perceived competence to hold a certain professional position would not be in any way diminished if he were black – if, that is, blacks with comparable competence have been hired to such positions, then the charge of

reciprocal higher-order political discrimination may be defeated. Second, ascertain whether there is any alternative property, conduct or manner, directed toward the same goal – of gaining access to some social advantage, that avoids or deflects the value conferred by the primary valued property. If there is not – if, for example, whether you are assertively ambitious or carefully diplomatic, intellectually prolific or intellectually fallow, you can do no wrong, then the charge of reciprocal higher-order political discrimination is prima facie justified.

Here it might be objected that the second test is inadequate to ascertain the existence of reciprocal higher-order political discrimination, since the explanation for why "you can do no wrong" may be, not that all such higher-order properties receive value from primary valued properties, but rather that all such higher-order properties are in any case irrelevant to judgments of a person's competence. However, remember that the second test applies specifically to properties directed toward the goal of gaining access to some social advantage. This includes not only properties irrelevant to the question of one's entitlement to that advantage, such as those pertaining to the manner or quality of one's self-promotion, but also properties directly relevant to that question, such as those pertaining to one's status, potential, training, experience, etc. The second test sifts out those cases in which irrelevant higher-order properties are made the basis for conferring the advantage, e.g. one's manner of self-promotion, and in which relevant higher-order properties are discounted as the bases for conferring the advantage, e.g. one's previous professional experience. In both kinds of cases, higher-order political discrimination is marked by the relaxation or modification of the criteria of competence for receiving the advantage, in order to accommodate the particular properties of the valuee.

Henceforth I shall take higher-order political discrimination to include reciprocal higher-order political discrimination. These two phenomena demonstrate that one need not be a blatant racist, sexist, anti-Semite, or homophobe – let us describe such an individual as a *simple first-order political discriminator* – in order to practice political discrimination. Higher-order political discrimination is given fullest expression indirectly, by implication, in seemingly unrelated tastes, preferences, and behavior.

4.3.3. Denial

So far I have used locutions like "seen as conferring value/disvalue on" and "by reason of its association with" to describe the relation between primary and higher-order disvalued or valued properties, without saying in any detail in what I take that relation to consist. It does *not* consist in the set of beliefs held by the higher-order political discriminator to the effect that

(A) (1) agent A has primary disvalued property P;
(2) agent A has *n*-ary property N; and
(3) P confers negative value on ... N.

(A) is faulty because of (3): Only the most perverse and unrepentant higher-order political discriminator would admit – even to herself – that it is P that confers negative value on N. On

the other hand, only the most absurdly consistent higher-order political discriminator would affirm the belief that, in virtue of (A.1) and (A.2),

(3') therefore N is of negative value, period.

This would be the plight of the higher-order political discriminator who, in virtue of his contempt for Isaac Asimov's intellectual prolificity, would feel compelled to abjure Balzac as well. Instead, (3) must be replaced by

(3") N, in the way in which it is borne by A, is of negative value.

(3") is better because it incorporates that locution that scrupled higher-order political discriminators are so reluctant or unable to further define: For the higher-order political discriminator, there is just something about *the way in which* a person dances rhythmically that is vulgar; something about the way in which a person manifests her intelligence that is glib, clever or sophistical; something about the way in which she attempts to gain access to social advantages that is unctuous or opportunistic. The higher-order political discriminator would vehemently reject the suggestion that this "something" might have anything to do with the person's race, gender, class, sexual orientation, or ethnic or religious affiliation. But in fact, it is precisely this primary disvalued property from which the blemish spreads. Let us then take the following set of beliefs

(B) (1) agent A has primary disvalued property P;
(2) agent A has *n*-ary property N; and
(3") N, in the way in which it is borne by A, is of negative value,

plus the following stipulation

(4) For the higher-order political discriminator, A's possession of P is what in fact confers negative value on N

as characteristic of the typical, i.e. scrupulous higher-order political discriminator.

What makes higher-order political discriminators so scrupulous? What, that is, explains the higher-order political discriminator's tendency to suppress (B.4)? Part of the answer lies in the nature of first-order political discrimination. As we have seen, first-order political discrimination can be understood as a species of pseudorationality that relies heavily on the mechanisms of rationalization and dissociation. The perception of someone's race, gender, class, sexual orientation, ethnic or religious affiliation, etc. as a source of his disvalue or value is the consequence of applying value concepts like "person," "human being," "citizen," "member of the community," "rational and responsible agent," etc. too narrowly, to include only those individuals who have the primary valued property, and exclude those individuals who lack it. And similarly, dissociating Jews as subhuman, blacks as childlike, gays as perverts, working class people as animals, or women as irrational are ways of obscuring one's identification of these individuals as fully mature, responsible human beings, and thereby obscuring one's recognition of these individuals as full members of the community with which one identifies.[5]

Higher-order political discrimination then adds to this the pseudorational mechanism of *denial*, in which we suppress recognition of an anomalous thing or property altogether, in order

to preserve the internal consistency of our beliefs or theory about the world, ourselves, and other people. I have already argued that typically, higher-order political discriminators are likely to be first-order political discriminators as well; that is, they have the same prejudices that incline them to view individuals with the primary disvalued properties as inferior, not fully members of their moral community. The simple first-order political discriminator experiences no conflict in categorizing disvaluees as inferior beings to be suppressed and exploited. Therefore, she has no need to exercise denial, either of her own discriminatory responses or of the disvaluees' existence. By contrast, higher-order political discriminators must deny both, in order to preserve horizontal and vertical consistency over time. Because they are deeply affected, but not fully reformed, by arguments and experiences that suggest that political discrimination is unjust, both their own discriminatory responses and the objects of those responses are anathema to higher-order discriminators. Because they do not want to believe that their responses are politically discriminatory, they deny them altogether. The higher-order political discriminator may deny, for example, that the primary disvalued property in question is a disvalue at all, and yet helplessly deplore the "fact" that nevertheless there are no competent or worthy candidates bearing this property to be found; or hold any such candidate to a much higher standard of acceptance or performance than that he ordinarily applies, relative to which her secondary properties can be disparaged. He may denigrate her intelligence as cleverness; or ridicule her for working too hard when she exhibits energy and commitment to her work; or disparage her professional recognition as achieved through hustling or connections.

These discriminatory responses suggest that the higher-order political discriminator in fact categorizes such members of the disvalued group themselves in similarly demeaning terms with respect to their primary properties, but experiences a conflict of conscience about doing so. Faced with the conflict between first-order politically discriminatory habits of thought and the dictates of conscience, the higher-order political discriminator exercises denial, above all in order to avoid this conflict, by eradicating its source from awareness. The higher-order political discriminator often fails to acknowledge the very existence or presence of members of the disvalued groups, in order to circumvent his own, first-order politically discriminatory responses to them.[6] For instance, he may ignore or fail to acknowledge a disvaluee's contribution to a general discussion, or respond to that contribution as though someone else had made it. Or he may relegate a disvaluee to marginal or peripheral tasks in a professional setting. Or he may simply ignore the disvaluee altogether, avoiding all social interaction not strictly required by social or institutional obligations. In behaving in this fashion, the higher-order political discriminator does not give vent to any sort of malevolent impulse. His aim is not to insult or injure the disvaluee. Rather, his aim is to avoid the painfully conflicting feelings – of disgust or contempt on the one hand, and the pangs of conscience on the other – that acknowledgement of the disvaluee provokes.[7]

Denial of a person's presence as a way of avoiding conflicting feelings about them is fairly common. A very handsome man may be the object of denial, when others' feelings of

attraction to him conflict with their conviction that these feelings are inappropriate; a very fortunate or charismatic person may be the object of denial, when others' feelings of envy or resentment conflict with a similar conviction. Or a homely person may be the object of denial when others' feelings of repugnance conflict with their kindness or social good will. Higher-order political discrimination is most analogous to this last-described case.

When social or institutional obligations make denial of the disvaluee's presence impossible, denial of (at the very least) her primary disvalued property, and of its perceived disvalue, supplies a second-best resolution to this conflict of conscience: Denial of the disvaluee's primary disvalued property suppresses from awareness the discriminatory habits of thought elicited by it, hence similarly preserves horizontal and vertical consistency over time, by placating the requirements of conscience. Thus the higher-order political discriminator is guilty of an even greater failure of cognitive discrimination than that of the simple first-order political discriminator. For whereas the latter fails merely to perceive the disvaluee's personhood through her difference, the latter fails to perceive either her or her difference at all. This is why the higher-order political discriminator tends to suppress (B.4). Unfortunately, to suppress habits of thought from awareness is not to eradicate their influence, any more than to suppress the disvaluee's existence from awareness is to eradicate her influence. Higher-order political discrimination is characterized by that attitude in which a certain habit of thought, namely first-order political discrimination, poisons one's evaluations and behavior, whether one acknowledges this or not.

The higher-order political discriminator is inclined, moreover, not to acknowledge this, no matter how obviously incriminating his evaluations and behavior may be to a disinterested observer. For this would expose the painful conflict of conscience the higher-order political discriminator's behavior attempts to suppress. To acknowledge this conflict, in turn, would be to acknowledge the need to resolve it, i.e. the need to work through and overcome the first-order prejudices that gave rise to it. But it is precisely in virtue of those first-order prejudices themselves that such a project of self-improvement stands very low on the higher-order political discriminator's list of priorities. Unlike the resolution of Oedipal conflicts, emotional problems, tensions in one's personal relationships, or career dilemmas, coming to terms with one's prejudices and learning not to inflict them inadvertently on others just is not, in the last analysis, seen as terribly important by the higher-order political discriminator. But it would be wrong to interpret the higher-order political discriminator as concerned only with personal problems and not with social ones. Rather, the higher-order political discriminator belittles the importance of addressing a very specific *personal* problem. That is part of what makes him a political discriminator in the first place.

As I have painted it, then, higher-order political discrimination is peculiarly the sickness of thoughtful, well-intentioned and conscientious individuals who nevertheless have failed adequately to confront and work through their own prejudices, or who perhaps have been too quickly satisfied by their ability to marshall arguments on behalf of doing so. One implication of

characterizing higher-order political discrimination as a sickness rather than a fault is that higher-order political discriminators are, in the last analysis, not morally responsible for their behavior. This conclusion may seem unpalatable in many respects. Nevertheless, direct appeals to reason in higher-order political discriminators are unlikely to work, because their dogged pseudorationality is so inherently a part of the problem. Such individuals are being neither disingenuous nor hypocritical when they deny that a person's race, gender, class, sexual orientation, or ethnic or religious affiliation affects their judgment of her competence or worth. They vehemently insist that this is so, they want it to be so, and they genuinely believe it to be so. They are, nevertheless, mistaken. Their efforts to explain away each manifest expression of higher-order political discrimination on different and inconsistent grounds are unconvincing. And their behavior exhibits a degree of otherwise inexplicable arbitrariness and idiosyncrasy that severely strains our attempts to apply the principle of charity in making sense of it. Hence in order to understand the behavior of higher-order political discriminators, we must watch what they *do*, not what they *say*.

For example, these attitudes may find expression in an expectation of greater deference or genuflection from a member of the disvalued group. The simple first-order political discriminator expresses his anger at the violation of this expectation in certain familiar stereotypes: the "uppity nigger" whose refusal to behave subserviently is seen as impudence or disrespect; or the "Jewish-American Princess," whose assertiveness, presumption of self-worth, and expectation of attention and respect is seen as a sign of being spoiled, selfish, or imperious. But for the higher-order political discriminator, such anger is displaced into more subtle but similar reactions: Such an individual may just feel angered or personally affronted by a woman's presumption of equality – in personal, social, or intellectual status, or professional worth, or as a competitor for social or professional rewards; or unduly irritated by her failure to defer or back down in argument. She may be viewed as forward in conversation, when in fact she contributes no more and no less than anyone else; or stubborn, unresponsive, or impervious to well-intentioned criticisms, when in fact the only acceptable response to those criticisms, in the eyes of the higher-order political discriminator, would be for her to concur with them wholeheartedly and apologize for her dereliction.

Or, to take another example, the higher-order political discriminator may feel invaded or compromised by an African American's jocularity or willingness to trade friendly insults that one accepts as a matter of course from those considered to be one's peers. The African American may be viewed as overly familiar, insolent, or presumptuous. In all such cases, the disvaluee's behavior is seen as a presumption, not a right or an accepted practice. The view of the disvaluee's assumption of equality as a *pre*sumption may explain the higher-order political discriminator's otherwise inexplicable umbrage at being complemented by a disvaluee: An inferior is in no position to confer favors of any kind. Thus the higher-order political discriminator is tortured by the suspicion that he is somehow being ridiculed, or shown insufficient respect, or that the disvaluee's conduct bespeaks contempt.

In a compelling analysis of anger,[8] N. J. H. Dent suggests that anger is based ultimately on feelings of personal inferiority: These lead one to overestimate the importance of others' expressions of regard and esteem for one, which in turn multiplies the number of occasions upon which one feels slighted or demeaned when such expressions are not forthcoming, or of insufficient magnitude relative to one's importunate requirements. I argued in Volume I, Chapter II.2.4 that feelings of inferiority were a necessary constituent of a Humean self. The oversensitivity to being slighted that Dent describes is a natural concomitant. Dent argues that this oversensitivity in turn provokes in one the desire to rectify one's situation through retaliation, by lashing out at the offender. Dent's analysis by itself does not, I think, cover all cases of anger; nor does it explain the origins of simple first-order political discrimination. But it does provide insight into why higher-order political discriminators, like simple first-order political discriminators, are apt to become so angry, so often, at imagined slights from seemingly arrogant disvaluees. The more inferior one feels, the more expressions of esteem one requires. And the more inferior one perceives a disvaluee to be, the more elaborate the disvaluee's expression of esteem of one is required to be. Whereas a friendly nod from a perceived superior is sufficient to transport one to a state of grace, anything less than a full-length obeisance from a perceived inferior appears to be an insult. In the American Deep South up to the mid-1960s, for example, for an African American to meet the gaze of a European American was perceived as an offense; and for an African American man even to look at a European American woman was to invite lynching. Even now, African Americans are still expected to do rather too much grinning and shuffling compared to their European American counterparts, although the retaliatory sanctions for disobedience are now a bit more oblique. In all such cases, irascibility regularly directed at particular members of disvalued groups should not be dismissed as simply an idiosyncrasy of character, even if it is not intentionally directed at members of disvalued groups *as such*. It is, nevertheless, an overt expression of higher-order political discrimination.

A second, related example of behavior and judgments distorted by higher-order political discrimination is the treatment of disvaluees in a way that would constitute a clear insult or *faux pas*, if the person so treated were one of one's recognized peers. For example, a European American Gentile may privately make an anti-Semitic remark to an African American colleague, in a misguided effort to establish rapport, when such a remark would be seen as a serious social lapse even among other European American Gentiles. Or a heterosexual may make gratuitous disparaging remarks to a gay colleague about her work or job performance, of a sort designed to "cut her down to size" rather than provide constructive criticism. Or a man may make offensively personal remarks to a woman colleague about her physical appearance, personal life, or manner of dress, of a sort that would be highly inappropriate if they were made to another man. Or he might expect from a woman colleague extra forbearance for fits of temper, irresponsible conduct, or extraordinary professional demands that he would not from a man. The higher-order political discriminator, in other social contexts, may be acclaimed quite rightly as a "prince among men"; to disvaluees, however, he reveals himself as Mr. Hyde.

This often creates additional difficulties in identifying cases of higher-order political discrimination for what they are. Because a disvaluee's insight usually remains morally unrecognized by the surrounding community (thus violating inclusiveness criterion (3) of the preceding chapter), her testimony suffers a credibility problem at the outset. This problem is severely exacerbated if the testimony concerns a higher-order political discriminator whom others have every reason to regard as a saint. Under these circumstances, any charge of inconsistency – whether it comes from others and targets the disvaluee, or comes from the disvaluee and targets the higher-order political discriminator – is in the eye of the beholder. For higher-order political discriminators regard coarse, tasteless, or brutal behavior toward disvaluees as called forth by them and so warranted; hence as fully consistent with the most highly refined manners and courtly civility toward others. Yet unlike former President Lyndon Johnson, who conferred with his cabinet through an open bathroom door, while uninhibitedly and indiscreetly performing his morning ablutions, the higher-order political discriminator cannot be supposed to commit these boorish excesses with any offensive intent. Rather, he regards his response to a person's disvalued properties as socially innocuous; as an acceptable variation in social etiquette, keyed to the variations among the personality traits of the variety of individuals he encounters.

A third example of such distorted behavior is the implicit treatment of disvaluees as being obligated by different rules of conduct than those which govern oneself and those considered to be one's peers. One may apply different criteria of interpretation to the behavior of disvaluees: Whereas enigmatic behavior by valuees is excused, overlooked, or given the benefit of the doubt, similar behavior on the part of disvaluees is interpreted as proof of vice or malevolence. This interpretation motivates the higher-order political discriminator not only to avoid, but also to justify the avoidance of direct interaction with the disvaluee, and thus avoid the conflict of conscience described earlier. Or one may apply rules of honor, loyalty, and responsibility only to those considered to be one's peers, but may have no scruples about betraying the trust or confidentiality of a disvaluee, who is implicitly viewed as unentitled to such consideration. Alternately, one may hold disvaluees to far more stringent moral standards than the members of one's own community in fact practice among themselves. Any violation of these standards by the disvaluee then creates an irradicable moral blemish to which the valuees are invulnerable, by reason of their status as valuees. These cases express quite clearly the conviction that disvaluees just do not have quite that same status, hence are not to be subject to the same standards of treatment, as members of one's recognized community – at the same time that the higher-order political discriminator vehemently and in all honesty denies that any such discrimination is taking place. Indeed in all of these examples, the higher-order political discriminator may sincerely deny that the person's race, gender, sexual orientation, ethnic or religious affiliation, etc. arbitrarily influences his evaluations, when his behavior shows patently that they do.

4.3.4. Exacerbation

There are many forces that may exacerbate higher-order political discrimination and its social consequences. Among them are, first and foremost, complicitous institutional practices. As we have seen in Chapter X, individuals in positions of responsibility may rank their personal and social allegiances ahead of their professional obligation to protect disvaluees from the pernicious effects of higher-order political discrimination. Or they effectively reward it, by regularly interpreting instances of it as expressions of professional autonomy, and refusing in principle to scrutinize suspected instances of it, on the grounds that doing so would be unwarranted interference in an organization's internal affairs. Institutions whose internal equilibria depend on such complicity are bully systems in the sense defined in Chapter X; and the arguments offered there apply. These institutions often comply with the letter of anti-discriminatory policies, by hiring members of disvalued groups to temporary positions of high public visibility. Such individuals are usually in a small and powerless minority, and regularly create cognitive discomfort in the majority whose territory they invade. Predictably, such individuals are then forced out – through harassment, discrimination, retaliation, or "downsizing" – and replaced by other, equally competent but equally transient members of the same disvalued group. Hence that group's symbolic representation within the institution can be maintained, without infiltrating its entrenched system of political discrimination through permanent or seniority status. The paucity and transience of such representative individuals then enable those who benefit from a bully system to deceive themselves about the depth and seriousness of their higher-order political discrimination, and to indulge without interruption fantasies of the tolerance and generosity they would manifest, were there more such individuals among them. This is to abdicate responsibility for enforcing those anti-discriminatory policies to which such institutions publicly claim to be committed. But bully systems such as these do not thereby evade their moral accountability, because they knowingly and deliberately deceive both the public and the new recruits on whom their continued existence depends.

Second, there is the intellectual resourcefulness of the higher-order political discriminator: Someone who is in fact deeply invested in the disvaluational status of some primary property may always recruit some further, equally irrelevant property to explain her seemingly irrational judgment, and thus deflect the charge of higher-order political discrimination: It may be said, for example, that the disvalued property is not a person's race, gender, sexual orientation, class, ethnic or religious affiliation, etc., but rather his inability to "fit in," to "get along with others," or "be a team player". This is a particularly familiar and dependable response, because the evidence for ascribing this property may be materially coextensive with the evidence for disvaluing the primary property at issue: Since the disvaluee is in theory held to the same standards of conduct that govern others in the community, but in fact expected to conform to different ones, tailored to his disvalued status, his inability to "fit in" can be guaranteed at the outset.

Under these circumstances, the disvaluee may collaborate in the fiction of fair treatment through his own pseudorationality, if his personal investment in the theoretical standards of fair treatment is so great that he rationalizes, dissociates, or denies the facts of discrimination that poison his life. Indeed, the disvaluee may well be even more stubbornly invested in this theory than the higher-order political discriminator, to the extent that his self-respect depends on believing that his own experience confirms it. He may find it simply unthinkable to contemplate the possibility that the respect and good will he accords to his colleagues is being returned to him with fear and contempt. But I argued in Chapter IX that literal self-preservation requires that, although such ideals ultimately must die, they must not do so without a long and painful struggle. Awaiting the disvaluee at the end of that struggle is the clarity of perception and insight into bully systems conferred by his disvalued and theoretically anomalous status.

A third force that intensifies higher-order political discrimination are the repressive, pseudorational habits of rationalization, dissociation, and denial already discussed. Earlier I suggested that higher-order political discriminators were generally well-intentioned individuals who had failed to come to terms with their own prejudices. I also mentioned some possible reasons for this failure: avoidance of conflicts of conscience, feelings of personal inferiority, and first-order political discrimination among them. Another reason that should not be neglected is that higher-order political discriminators tend to rationalize, dissociate, or deny the very existence of higher-order political discrimination itself. They might claim, for example, that the phenomenon I have described is in truth perceptual sensitivity to subtle variations and qualities among individuals, all of which might be relevant to questions of value or competence in a sufficiently broad sense. Or they might agree that higher-order political discrimination exists, but dissociate it from their own motives and behavior, as an anomalous phenomenon that is too rare to merit further scrutiny. Or they might just flatly deny the existence of anything like what I have described as higher-order political discrimination, and deny as well the undeniably familiar instances of it that I have invoked to anchor the foregoing analysis. These tactics reinforce the tendencies of higher-order political discriminators to deny their own collusion in the practice of higher-order political discrimination, and to deny or minimize their need to come to terms with it. Higher-order political discriminators are adept at the tactics of pseudorationality because they have so much self-esteem to lose by modifying their beliefs. But we must not be taken in. For above all, higher-order political discriminators need to understand that no one is fooled by their tactics. With the aid of this understanding, they may someday learn to stop fooling themselves.

5. Corrigibility and Vertical Consistency

How might higher-order political discriminators come to such an understanding? Are they even capable of achieving self-awareness of the pseudorational tactics that buttress their political discrimination, and the deep-seated xenophobia that fuels it? Recall the Kantian thesis on which Chapter II.4 was based, that if a perception fails to conform to the categories of thought that unify and structure the self, it cannot be experienced by that self at all. Applying this general

thesis in Chapter VIII, I argued that if we cannot make sense of the data of third-person moral anomaly in terms of the familiar concepts that structure our experience, we cannot register it as one of them. I also distinguished, in Section 2 above, between the innate idea of personhood as a hard-wired – or, in Kant's terminology, transcendental – concept; and our empirical, contextually determined conception of people. And in Section 3.1 above, I argued that sometimes, our empirical conceptions of other people are so limited that if an individual is unfamiliar enough, we may be incapable of discerning her personhood through the theoretically anomalous and threatening manifestations of her empirical personality. This suggests two ways in which the cognitive failures that underlie higher-order political discrimination might function:

> (A) A higher-order political discriminator might regard someone as fully a person if and only if she also recognizes him as falling within her familiar conception of people; or
>
> (B) She might recognize him as a person even if he violates her limited and familiar conception of people.

(A) is the dogmatist described in Chapter VII.4.4, who conceives her experiences as hers if and only if she conceives them as instantiating her favored theory – in this case, her favored theory about people in general. If (A) describes my cognitive failings, then an anomalous other who violates my limited conception of people thereby violates my transcendental conception of personhood as well. I am then strongly disposed to regard such a being as a thing, or as an animal, or as subhuman or unnatural or unholy, or in any of the other similar ways by which we demonize others in order to rationalize our mistreatment of them. On the proposed Kantian conception, this cognitive disposition has a deep cause. We have already seen in Chapter II.4 that the concept of personhood is at best an instantiation of the transcendental substance-property relational category. Since my transcendental concept of personhood is not equivalent to the transcendental concept of a thing or substance in general, my failure to recognize the other's personhood does not imply a failure to recognize her as an object with properties altogether. I may recognize another who is anomalous with respect to my concept of personhood as consistent with my concept of objects in general. However, if the other must conform to my limited conception of people in order to conform to my concept of personhood – i. e. if something is a person for me if and only if it falls under my empirical conception of people, but does not, then from my perspective, an object is all that she can ever be. In this case, my xenophobia in general and political discrimination more specifically is a hard-wired cognitive disposition that is impervious to empirical modification.

But suppose instead that the higher-order discriminator's cognitive malfunctions are better described by (B). (B) is the case in which an otherwise unfamiliar object – in this case, an anomalous subject – is subsumed under the highest-order concept of the self-consciousness property, i.e. as an experience she has, even if there are few lower-order concepts in her empirical arsenal that would render it familiar to her. So (B) describes a xenophobe whose cognitive condition satisfies vertical consistency to some degree. (B) leaves open the possibility that a

person might have an empirically limited conception of people yet fail to be a xenophobe, just in case she acknowledges as a matter of principle that there must be other ways to do things and other ways to live besides those with which she is familiar; and just in case she is able to put this principle into practice when confronted by some of them. This is the case described in Section 4 and 4.1, of the individual who commits cognitive errors 3.1 – 3.3, but has no personal investment in the defective empirical conception that results.

(B) thus leaves open the possibility that one could be a xenophobe in the sense discussed in Section 4 and 4.1, yet be corrigible in one's xenophobia. For (B) acknowledges the possibility that even though the xenophobe equates her limited conception of people with her transcendental concept of personhood, someone might conform to her transcendental concept of personhood without conforming to her empirical conception of people. That is, in this case it is cognitively possible to introduce into her range of conscious experience a new object the behavior of which satisfies the rule and order of rationality even though it fails to satisfy her honorific stereotype of personhood. And it is possible for her to recognize in this conceptually anomalous behavior the rule and order of rationality, and so the personhood of another who nevertheless violates that honorific stereotype.

Since recognition of the existence of such a theoretical anomaly constitutes a counterexample to her honorific stereotype of personhood, the xenophobe has two options, according to (B). Either she may, through the mechanisms of pseudorationality, seek some strategy for explaining this anomaly away; or else she may revise her stereotypic and limited conception of people in order to accommodate it. Thus (B) suggests that it is in theory possible for the xenophobe to reformulate and reform that conception in light of new data that disconfirms it, and so to bring her reciprocal stereotypes closer to open-ended inductive generalizations.

Of course whether or not this occurs, and the extent to which it occurs, depends on the virulence of her xenophobia; and this, in turn, on the extent of her personal investment in her honorific, stereotypical self-conception. But to the extent that (B) is correct, to the extent that one can discern the personhood of someone who violates one's limited conception of people, pseudorational dismissal of the stranger as a person is not a viable option. By hypothesis the properties that constitute her identity as a person cannot be denied. Attempts to dissociate them, i.e. to dismiss them as insignificant, alien or without value have unacceptable implications for one's own which similarly must be pseudorationalized out of the picture. Moreover, attempts to rationalize them as flukes or mutations or illusions or exceptions to a rule undermine the universality of the rule itself. As in all such cases, pseudorationality does not, in fact, preserve the rational coherence of the self, but only the appearance of coherence in one's self-conception, by temporarily dismissing the theoretical anomaly that threatens it. In the event that a xenophobe is confronted with such a phenomenon, xenophobia conflicts with the requirements of literal self-preservation and finally must be sacrificed to it. So finally, the only way for this

type of xenophobe to insure literal self-preservation against the intrusion of an anomalous person is to revise her reciprocal stereotypes of herself and others accordingly so as to integrate her.

6. Kant on the Xenophilia in Vertical Consistency

There is evidence in the text of the first *Critique* that supports 5.(B) as Kant's preferred alternative as well. These are in those introductory, explicative sections of the Dialectic, in which Kant maintains that it is in the very nature of transcendent concepts of reason to have a breadth of scope that surpasses any set or series of empirical experiences we may have; indeed, to provide the simplest unifying principle for all of them and more. Thus, for example, he tells us that "the principle peculiar to reason in general, in its logical use, is: to find for the conditioned cognitions of the understanding the unconditioned whereby its unity is brought to completion." (1C, A 307/B 364) By the "conditioned," Kant means those experiences and rules that depend on an inferential relation to other, more inclusive principles that explain them. And by the "unconditioned," Kant means those principles, concepts or ideas of reason that are not themselves dependent on any further ones but rather provide the explanation of all of them. What he is saying here is that rationality works interrogatively for us: Given some datum of experience we understand, we reflexively seek to enlarge our understanding by searching for further data by which to explain it.

Kant then goes on to say in the same passage that this logical principle becomes a transcendent one through our assumption that if dependent explanatory rules and experiences are given, then the whole series of them, ordered in relations of subsumption of the sort that characterize a covering-law theory, must be given as well; and that this series is not itself dependent on any further explanatory principles; we have already seen in Chapter V.5.2 that Kant's substantive moral theory satisfies this condition as well. Kant's point is that we assume that any limited explanation of experience we have is merely part of a series of such explanations that increase in generality and inclusiveness, up to a maximally inclusive explanation of all of them, just as the criterion of vertical consistency requires. Thus, he argues, we regard each such partial experience of the world we have as one among many, all of which are unified by some higher-level theory. And later he says that

> [t]he transcendental concept of reason is none other than that of proceeding from a totality of conditions to a given conditioned. Now since only the unconditioned makes the totality of conditions possible, and conversely the totality of the conditions is itself always unconditioned; so a pure concept of reason in general can be explained through the concept of the unconditioned, so far as it contains a basis of the synthesis of the condition. (1C, A 322/B 379) ... concepts of pure reason ... view all experiential knowledge as determined through an absolute totality of conditions. (1C, A 327/B 384; also see A 311/B 368, B 383-385, A 409, A 509)

What he means is that we regard any particular phenomenon as embedded in a systematically unified series of such phenomena, such that if we can explain some partial series of that kind,

then there is an entire series of which that partial series is a part that we can also explain; and such that that more inclusive explanation explains everything there is about the phenomenon to explain. So Kant is saying that built into the canons of rationality that structure our experience is an inherent disposition to seek out all the phenomena that demand an inclusive explanation, and to test its inclusiveness against the range of phenomena we find.

These remarks support 5.(B) because they imply that the innate cognitive concepts that structure and unify our experience invariably, necessarily outstrip our empirical conceptions of it. Kant is saying that it is in the nature of our cognitive limitations – i.e. that we can only have knowledge of sense-based experience – that the explanatory scope of the innate concepts that structure and unify it necessarily exceeds that sensory basis itself. This means that we view any experience in implicit relation to other possible experiences of its kind, and finally in relation to some systematic explanation that makes sense of all of them; in the end, he thinks, we are so constructed as to require vertical consistency all the way up. However, no single experience, or series of experiences, can ultimately satisfy our appetite for conceptual completeness, because the scope of the higher-level concepts we invoke to explain them necessarily outstrips the limited number of those experiences themselves. There will always be a lack of fit between our innate rational capacity and the empirical theories it generates, because they will always appear limited in scope in a way our innate capacity for explanation itself does not. So no matter how much sensory data we accumulate in support of our empirical theories of ourselves or the world, we are so constructed intellectually as to be disposed to feel somewhat dissatisfied, inquisitive, restless about whether there might not be more to explain, and to search further for whatever our search turns up.[9]

But this means that we are disposed reflexively to regard anomalous data as more than mere threats to the integrity of our conceptions of the world and ourselves, for the disposition to inquire further and to seek a more inclusive explanation of experience remains, even when literal self-preservation has been achieved. We also are disposed to regard those data as irresistible cognitive challenges to the scope of our conceptions, and as provocations to reformulate them so as to increase their explanatory reach. Because, according to Kant, we are always seeking the final data needed to complete the series of experiences our conceptions are formulated conclusively to explain, it could even be said that we are disposed actively to welcome anomalies, as tests of the adequacy of the conceptions we have already formulated.

When applied specifically to the transcendent idea of personhood, this disposition to welcome theoretical anomaly as a means of extending our understanding amounts to a sort of xenophilia, a positive valuation of human difference as intrinsically interesting and therefore worthy of regard, and a disvaluation of conformity to one's honorific stereotypes as intrinsically uninteresting. It dismantles the assumption that there is any cause for self-congratulation or self-esteem in conforming to any stereotype at all, and represents anomalous others as opportunities for psychological growth rather than mere threats to psychological integrity. It implies an attitude of inquiry and curiosity rather than fear or suspicion, of receptivity rather than resistance

toward others; and a belief that there is everything to be gained, and nothing to be protected, from exploration of another person's singularity.[10] We often see this belief expressed in the behavior of very young children, who touch, poke, prod, probe and question one without inhibition, as though in knowledge of another there were nothing to fear. What they are lacking, it seems, is contingent empirical evidence to the contrary.

7. Xenophilia and Aesthetic Anomaly

In those of us for whom 5.(B) is the right interpretation of our cognitive attitude toward anomalous others, contemporary art offers a training ground for cultivating the xenophilic disposition to inquiry by which we may temper the refined xenophobic excesses of higher-order political discrimination. I do not mean to suggest that works of art are capable of curing higher-order political discrimination. As we have seen, higher-order political discrimination is supervenient on first-order political discrimination; and first-order political discriminators are ashamed, not of their political discrimination, but of themselves as inadequate to the honorific stereotypes they reciprocally impose on themselves. In so far as a higher-order political discriminator retains a personal investment in that honorific stereotype, she will be unpersuaded by its deleterious effects on others to renounce it. This means that it is not just her cognitive habits that are in need of reform, but her more central conception of herself. This is a task for social reconditioning or psychotherapy, not art. Nevertheless, art has an important role to play in intensifying a viewer's self-awareness of these matters. Art can highlight pseudorational failures of cognitive discrimination as themselves objects of aesthetic examination; and it can heighten a viewer's level of cognitive sensitivity to a wide range of complex situations, of which political discrimination is only one.

In the contemporary setting, galleries and museums announce themselves to the public as arenas in which cognitive alertness is required, and in which the viewer's capacity to understand and situate an anomalous object in its singularly appropriate context will be tested. In earlier historical periods, galleries and museums had different roles: pedagogical or inspirational, for example. But in this one, their primary role, and the role of the art works they exhibit, is to challenge the limitations of the viewer's conceptual scheme – his presuppositions about reality, the human condition, and social and personal relationships, as well as his presuppositions about what art is and what an exhibition space is supposed to do. By introducing into a specialized cognitive context singular objects that defy easy categorization, galleries and museums signal themselves to their audience as purveyors of heightened awareness through the objects and artifacts they display. Generated by a culture that values innovation for its own sake as well as for its ability to create its own market, these contemporary artifacts function primarily to provoke or stimulate in the viewer more flexible and inclusive conceptualizations of reality that can encompass them. In this sense, contemporary art offers a deliberate and paradigmatic experience of theoretical anomaly. It provides one the opportunity to reorganize one’s favored theories, the self-conception with which they are intertwined, and

therefore the conceptual structure of the self in order to accommodate it; and to test and develop one's capacity for cognitive discrimination in order to grasp it.

Some works of art satisfy this desideratum better than others. Some choose instead to reaffirm traditional values, or the social and political *status quo*, or prevailing comfortable convictions and perceptions of human nature. But since Impressionism and perhaps before, but most explicitly since Duchamp, the most significant works of art in the Western tradition[11] have taken seriously the challenge of heightened cognitive discrimination, i.e. the challenge to compel the viewer to see what she did not see before, and to add these anomalous, newly discovered properties of objects and events to her permanent cognitive repertoire. Many contemporary artists take seriously their responsibility to question and extend the limits of knowledge by offering anomalous objects, innovative in form, content, or both, as an antidote to provincial and conventional habits of thought.

Minimal Art of the 1960s offers a particularly compelling example of this. For the first time in the history of Modernism, artists were taken seriously as critics and theorists of contemporary art. And what many Minimal artists explicitly averred in their writings was that no such theory was adequate to an understanding of the work; that the point of presenting geometrically, materially and formally reductive objects was to draw the viewer's attention away from extrinsic associations and toward the specificity and materiality of the particular object itself. In its aesthetic strategies, Minimalism repudiated the imposition of abstract theory – psychoanalytic, social, or aesthetic – as cognitively inadequate to a full comprehension of the work. Instead it emphasized the uniqueness, singularity, and indexical immediacy of the art object itself. The category of art itself functioned as a catch-all term signifying the object's inherent resistance to extrinsic conceptualization, and so its aesthetic interest as an otherwise anomalous entity in its own right. This stance itself was, of course, a theoretical one. But Minimalism differed from earlier theoretical stances in stipulating the properties of the specific object in question as the origin and locus of theorizing about it. It embedded the object in an abstract symbol system of its own making.

Conceptual and Performance art of the late 1960s and early 1970s extended this strategy further, by subordinating the medium in which the work was realized to the concepts it embodied or explored. It was even more clearly the intrinsic meaning of the work, and not the cognitive preconceptions the viewer brought to it, that dictated its appropriate conceptualization. In subordinating medium to concept, Conceptualism not only reaffirmed the conceptual fluidity and inclusiveness of art, as originally introduced by Duchamp's urinal. It also opened the door to the use of any medium, event or object deemed appropriate to the particular concepts the artist chose to explore. Thus Conceptualism repudiated all remaining traditional restrictions on content and subject matter as well as on medium. And in so doing, it created the possibility of seeing any object as a theoretical anomaly relative to the conceptual scheme within which it was conventionally embedded. Any such object became a potential locus of original conceptual

investigation, and all such objects became potential threats to the conceptual unity of a rigidly or provincially structured self.

Under these circumstances, the gallery or museum as a site of cognitive provocation has become clear. Beyond a few extremely vague and uninformative terms of classification, such as "installation art," "performance art," "object art," etc., there are no longer any expectations or preconceptions a viewer may legitimately bring to such work regarding what kind of viewing experience is in store – except that he will be required to discriminate cognitively a variety of elements, and fashion for himself a coherent interpretation of the experience that at the same time respects the intrinsic conceptual integrity of the work. A viewer of contemporary art must be prepared for media that include foodstuffs, bodily fluids, chemical compounds, and industrial materials, as well as traditional art media; and for content that may be highly autobiographical, social, sexual, political, or philosophical, as well as realistic or abstract. No viewer who insists on maintaining excessively rigid, provincial, or philistine views about art will survive in the contemporary art world for very long.

Thus the contemporary art-going public is self-selected primarily to consist, not in a specialized educational and economic elite (as though there were no working-class artists, self-made millionaire collectors, or scholarship students among the art critics); but rather in those individuals who are psychologically prepared to engage in the hard work of cognitive discrimination in general. For all of the above reasons, the contemporary art-going public is expected to be more than ordinarily receptive to the conceptual challenge presented by theoretically anomalous objects or properties in general, and, *a fortiori*, by theoretically anomalous persons in particular. The arena of contemporary art, then, invites examination and evaluation with respect to the goals of addressing the cognitive failures of xenophobia and redressing the moral failure to satisfy the proposed criteria of inclusiveness.

Now return to the plight of the higher-order political discriminator, taken in by her own pseudorational attempts to eradicate awareness of her xenophobic attitudes and behavior. With its latitude in the use of media, content, and subject matter, contemporary art may offer a variety of approaches for reducing this cognitive disingenuity and enhancing self-awareness. Take, for example, *mimesis*: A work of art may incorporate into its subject matter these very pseudorationalizations as an ironic commentary or distancing device. These pseudorationalizations not only impose politically discriminatory stereotyping on others. They are themselves stereotypical reactions, conditioned habitual responses that are part of a behavioral repertoire as limited as that which the political discriminator imposes on anomalous others. Indeed, they embody such stereotypes even as they express them. It is in the nature of deeply instilled habits of thought and action to seem, not only deeply private and individualized; but also fixed, natural, and part of the objective order of things – so much so that voluntarily bringing them to light as objects of self-conscious scrutiny on one's own is exceedingly difficult. One scarcely knows what to question or scrutinize. But hearing or seeing them echoed back to one by an impersonal art object can make it clear to one that these phrases or habits of reasoning

are not uniquely one's own, but rather crude and common slogans that short-circuit the hard work of self-scrutiny. Thus mimesis can be an effective way of distancing oneself from such pseudorational slogans, and of illuminating their stereotypical character and function. By demonstrating their indiscriminate and simplistic application to a range of circumstances that clearly demand greater sensitivity to specifics, such a work can encourage greater cognitive discrimination of particular persons and circumstances for what they are.

A second device that may be useful as an antidote to higher-order political discrimination is *confrontation*: As we have seen, a higher-order political discriminator escapes from the meaning of his behavior into a thicket of abstract pseudorational theorizing that detaches him from the actual personal and social consequences of his actions. Because he denies the existence of the object of his higher-order political discrimination, in addition to his own responses to it, the higher-order political discriminator often lacks a sense of the hurtfulness of his behavior, or of the harmfulness of its consequences for others. An art object that confronts a higher-order political discriminator with the human repercussions of these consequences can help restore to the higher-order political discriminator a sense of reality, and a sense of cognitive responsibility for the human effects of his unreflective stereotyping of anomalous others. Moreover, a confrontative art object can draw the higher-order political discriminator's attention away from the abstract realm of theoretical obfuscation, and back to the reality of his actual circumstances at the moment. It can help resituate him in the indexical present of his immediate, one-to-one relation to the object and the issues it embodies.

Finally, consider the strategy of *naming*: We have seen that pseudorationality for the higher-order discriminator consists in the construction of an elaborate edifice of euphemisms designed to obscure from herself and others the true meaning of her attitudes, actions, and policies toward others, and of the painful social realities to which her behavior in fact responds. This willed unconsciousness can be penetrated by concepts and symbols that speak plainly to the ugly realities these euphemisms conceal. An art object that draws the viewer's attention to these realities, and leaves no room for ambiguity in their identification, can be an assaultive and disturbing experience. It blocks escape into abstract speculation concerning the denotations and connotations of the terms or symbols deployed as referents, and may reinforce the vividness and objectivity of the realities brought forward through confrontation, with the legitimating imprimatur of linguistic or representational recognition. At the same time, through repetition and repeated viewing, it can help accustom the higher-order political discriminator to the existence of these realities, and conceptually defuse them to psychologically manageable proportions.

Of course each of these strategies, as well as many others I have not mentioned, can be deployed outside the contemporary art context as well as within it: in psychotherapy, encounter groups, or organizational training sessions, for example. But one benefit of utilizing art objects in this role is that, unlike psychotherapists, group leaders, or other human subjects, an art object can elicit different reactions from different viewers while maintaining exactly the same

phenomenological presence to all of them. It does not itself react personally to any particular viewer, or differently to one viewer than it does to another, or alter its presentational aspect to suit the tastes or dispositions of particular viewers. Because the logic of its internal structure and external appearance depends on its personal history and interactive relationship with the artist rather than with the viewer, its final form is fixed and immutable relative to any particular viewer in a way other human subjects cannot be. Thus a viewer's relation to an art object can be both direct and individual on the one hand, and impersonal on the other.

The impersonality, impenetrability, and inherent internal equilibrium of an art object can be a distinct advantage in attacking political discrimination through the cultivation of cognitive discrimination. A human subject who deploys these strategies in other interpersonal contexts is vulnerable to criticism by a participant who feels that the leader, trainer or therapist is "reacting personally" to him: who just doesn't like him, is personally attacking him, manipulating him, or projecting her own problems onto him. And in this type of situation, such criticisms may be justified. But in an art context, they cannot be. For unlike human subjects, an art object cannot have reactions to, intentions toward or designs of any kind on a viewer; and *a fortiori*, cannot have personal reactions, intentions or designs on any particular viewer. So although it may happen that a particularly insecure or provincial art viewer initially may feel moved to accuse the work of art of manipulating her, ridiculing her, trying to pull the wool over her eyes, guilt-tripping her, attacking her, etc., it will not require too much reflection on the viewer's part to conclude, finally, that this is not the kind of thing an art object, unlike a human subject, has the capacity to do. Nor will it require much more reflection on the viewer's part to conclude that, if she does indeed feel that the work is doing these things to her, these feelings can only be the result of magical thinking and personal projection of her own emotions onto the object; and that this response itself is worth her scrutiny. An important benefit of utilizing art objects to combat higher-order political discrimination, then, is that they enable the viewer to discriminate cognitively between what she sees and what she is.

Is there a difference between fine art and commercial art in this respect? Is the latter not clearly manipulative in intent? Not if we distinguish, in the case of art as well as of advertising, between the creator's intentions in producing the work, and its psychological effects on its viewers. Like advertisers, artists of course have intentions in producing a particular work. Typically, an advertiser's intention in producing a commercial is to get the consumer to buy the product, whereas an artist's intention in producing a work of art may be to get the viewer to reflect on his political or aesthetic attitudes. In both cases, these intentions can be distinguished from the psychological effects of the work on its recipient. An advertiser who pairs a beautiful woman with a certain make of car in order to get consumers to buy that make of car may intend to enhance the appeal of that make of car to consumers. That a particular consumer comes to hate his wife because he has a different make of car is not necessarily part of the advertiser's intention. Similarly, an artist who pairs depiction of the homeless with standard stereotypical rationalizations for ignoring them may intend to get viewers to reflect on their economic

priorities. That a particular viewer feel guilt-stricken because she has been making contributions to her alma mater instead of to the homeless is not necessarily part of the artist's intention. Any individual who engages in an act of communication of any kind intends to have an effect on his audience, at least minimally that it understand him. This does not imply that he intends the actual effect on his audience his communication has. A consumer as well as an art viewer may examine their reactions to a commercial and a work of art respectively, in order self-consciously to discern and differentiate their personal areas of vulnerability or uncertainty from the intended act of impersonal communication the object represents.

8. Xenophobia, Alienation and the Primacy of Principle

In *Alice Through the Looking Glass*, Lewis Carroll describes Alice as walking through the forest of things with no names. Because she has forgotten the name of everything, she fails to remember when things are so different and strange that she is supposed to be afraid of them. She encounters a fawn that similarly does not remember that Alice is a human being and that the fawn is supposed to be afraid of her. So they walk together through the forest, clasped arm in arm. When they come to the end of the forest, they remember that they are human being and animal respectively, and spring apart, terrified. Carroll's idea is that were we not confused by interposing classificatory terms, categories and concepts between ourselves and others, we would have the same, trusting closeness that Alice had with the fawn while they were in the forest. As long as we can forebear *labeling* one another, Carroll seems to suggest, we shall all get along just fine.

Carroll's suggestion is elaborated in Bernard Williams' concept of moral alienation, discussed in Volume I, Chapter VIII.3.2. As we saw there, Williams' argument is that moral alienation occurs when we interpose abstract concepts and principles of moral obligation between ourselves and other people, or between ourselves and those plans and projects that, he says, are most centrally definitive of who we are. The "one thought too many" is, in Williams' view, that which turns healthy personal interactions based on spontaneous mutual attachment into policy-driven formal transactions based on moral protocol. So Carroll's and Williams' views suggest that xenophobia and moral alienation go hand in hand: Both are engendered by "labeling;" by conceiving of others in abstract and general terms. And both can be defeated by foregoing the need to classify and categorize others in such terms, instead appreciating them for the uniquely complex and singular subjects they really are.

Now in this project, I have gone to some length to defend the primacy of abstract and general concepts and principles in the structure of the self. So I am not convinced that things would work out the way Carroll and Williams think they would. Without concepts and principles under which others' concrete particularity could be subsumed and rendered rationally intelligible, other people would be strange and cryptic entities whose behavior we would have to study in order to figure out how to use them to get what we needed. In essence we would treat other people as we ordinarily treat animals – as dinner, fur coats, glue, drugs, pet food, etc. This

would be a paradigm case of the egocentric and narrowly concrete perception of reality that would be left to us without the modal imagination that rational intellection supplies. Concepts and principles are absolutely central and crucial in the structure of the self, because they enable us to render rationally intelligible inherently enigmatic concrete particulars, and thus extend their meaning and significance for us, as well as our modal imagination of the interiors their enigmatic façades conceal, beyond the indexical present of immediate awareness. This is one reason why abstract principles can rationally motivate individuals to sacrifice personal projects or relationships for their sake.

Yet it is inescapably true that any such concept or principle we apply to any concrete particular, particularly human ones, is necessarily crude, relative to the unique singularity of the thing we apply it to. It is even harder to capture a person in concepts and principles than it is to conceptually capture any other concrete particular, even though without concepts and principles we could apprehend nothing at all. Because in fact each one of us is completely and utterly different from everyone else, no rule-governed term or concept, or conjunction of such, can be fully adequate to anyone's singular and complex reality. So each one of us violates as a matter of course the assumptions, expectations, and theories that others bring to bear on their experience of us. Each one of us is the conceptual anomaly we fear to find in others.

The resulting sense of anxiety, irritation, even panic at being thwarted in our attempt at epistemological control of another emerges with particular force when a person behaves or presents herself in ways that are not familiar or comfortable to us. This is the locus of the xenophobic impulse: that moment when another's unfamiliar appearance or behavior begins to violate our familiar presuppositions about her. That moment occurs with far greater frequency between individuals in close relationships than between strangers or groups, often with comparably destructive or even lethal results.[12] At one end of the spectrum, the limiting case of xenophobia is to be found when another person – an acquaintance, friend or loved one – starts to move into our psychological orbit. The closer the person comes, the more his strangeness and singularity begin to surface, and the more threatened we feel. Then the more we experience the need for cognitive control and the more we feel invaded by his violation of our space, our privacy, the boundaries of our interiority. At the other end of the spectrum, the more familiar limiting case of xenophobia is that of physical violence, rape, genocide, or territorial invasion, where the felt need is for physical control and the boundaries being invaded are physical rather than psychological. Both extremes and the large range of variations in between manifest the same xenophobic response. In the end, the personal really is, as is often said, political; and the quality of political discrimination is the same in all of them.

It is because the anxieties, conflicts and misunderstandings inherent in close interpersonal relationships are of a piece and psychologically continuous with the anxieties, conflicts and misunderstandings inherent in macroscopic political discrimination that abstractions such as nation, race, ethnicity, sexual identity, or religion can divide friends, couples, colleagues, co-workers and fellow citizens, and turn them into enemies overnight. The relative

crudeness and inadequacy of abstract and general concepts and principles to capture another's singularity diminishes neither their importance in the structure of the self nor their motivational efficacy under the right circumstances. The very same cognitive disposition that engages our interest in another and motivates us to learn more about her also restricts our capacity to know her; and motivates us to pseudorationalize, in distorted concepts and principles ranging from the benignly to the lethally ignorant, the theoretical anomaly the other represents.

Indeed, the very same concept or principle – for example, the liberation of one's country – can be a source of moral heroism on the one hand; and of moral alienation, personal betrayal or xenophobia on the other. The very same concepts and categories that structure the self and make an agent's experience coherent and meaningful are those which can turn friends into ideological enemies; or make the justified moral demand that friendship be sacrificed to the demands of principle; or prompt fear and hatred toward an anomalous other who appears to threaten or violate them. No particular type of moral or political theory can be hailed or faulted for this, nor is any particular moral or political content especially susceptible to it. It is rather our hard-wired cognitive apparatus that is the culprit. The deplorable facts that deep personal attachments can be flimsy in the face of theoretical provincialism or political ideology, and that even the most inspiring of moral principles is vulnerable to rigidity and provincial constriction in its application to actual moral agents, are the flip side of the sometimes salutary facts that matters of abstract principle of any kind can validly come between people or before profit; that one can validly choose to sacrifice a love relationship or one's family or one's career opportunities for the sake of moral commitment. All are by-products of the necessary conditions of unified experience on which this discussed has focused. So the problems of xenophobia and of moral alienation on the one hand, and respect for principle and moral integrity on the other, are two sides of the same coin. The coin is the inherently inscrutable nature of concrete particulars, and the challenge they present to the enterprise of rational intelligibility.

There are several familiar ways in which we commonly handle our xenophobic response when another seriously trespasses on the outermost boundaries – whether merely psychological, or also physical – of our theory-laden conception of ourselves, others and the world. First, we may fight. We may try to bully or manipulate the other into submitting to our preconceptions using verbal, emotional or physical coercion. I have detailed some of these tactics in the above discussion. A second familiar response is flight. Here we may simply pack up and shut down the dialogue, transaction or relationship. A third well-known response is to abandon the goal of meeting the requirements of inclusiveness outlined in Chapter X, and instead enter into an unspoken mutual agreement to fulfill each other's stereotypes. For example, one spouse can be the brainy wife and the other can be the caring, compassionate husband. Or one partner can be the hard-driving realist, and the other the sensitive idealist. Or one colleague can be the administrative whiz, another the charismatic guru, a third the unsocialized genius. Or one group can be the brash but virile source of military strength, the other the prudent but decadent source of civilization. Or one friend can be the wise, long-suffering martyr, whose role it is to inspire

and instruct the innocent, protected, lethally naïve and spiritually bereft other in how to behave toward members of the group – women, African Americans, Jews, Arabs, the working class – the first represents. Of course when we move to transcend these stereotypes and probe deeper complexities, we violate the comfortable psychological and cognitive boundaries they cement; and again call forth the xenophobic anxieties they were designed to placate. The moral of this story is that as crucial and central to the structure of the self as rationality is, it can take us only so far in tempering xenophobia, whether between individuals or among groups.

In the two preceding sections I have suggested some ways in which we might gradually redirect our powers of cognitive discrimination of others from the xenophobic to the xenophilic – from the frightened, dogmatic and rigidly defensive to the curious, interested and receptive – so as to strengthen the proper functioning of theoretical reason, by practicing on contemporary works of art. These suggestions were motivated by the conviction that good intentions of moral inclusiveness are not enough, and that no one can bootstrap herself out of xenophobia merely by willing it to disappear. The ultimate objective of such exercises, of course, would be concerted and prolonged application of the cognitive techniques learned there to those other people whom we reflexively exclude from full moral personhood. The proximate instrumental means to this objective would be to neither rationalize those techniques out of serious consideration, as frivolous mind-games to be indulged by members of the moneyed leisure classes; nor dissociate them as inapplicable to the serious world of morality and global politics; nor deny their efficacy even in those few and modest cases in which they actually seem to work. Whether we can muster the intellectual spine to forego any of these temptations of pseudorationality remains in question. Perhaps its answer depends on the strength of our recognition of ourselves in the capacities of transpersonal rationality, and the extent to which we are ennobled and transformed by discovering ourselves in it.

Endnotes to Chapter XI

[1]Thomas Kuhn, *The Structure of Scientific Revolutions* (Chicago: University of Chicago, 1971), Chapters VI-VIII.

[2]This is the main thesis of Professor Deborah Tannen's fascinating *You Just Don't Understand: Women and Men in Conversation* (New York: William Morrow and Co., Inc., 1990), a popularization of her research in linguistics on gender differences in language use.

[3] Here see Terrence Real, *I Don't Want to Talk About It: Overcoming the Secret Legacy of Male Depression* (New York: Scribner, 1997).

[4]I am indebted to Rüdiger Bittner for pressing this question in discussion.

[5]The irony in the case of racism is that there is a substantial literature in biology and the social sciences that indicates that almost all purportedly white Americans have between five and eighty percent black ancestry – hence are, according this country's entrenched "just one trace" convention of racial classification, black. For only a very small selection of the research that has emerged on this topic, see F. James Davis, *Who Is Black?* (University Park: Pennsylvania State University Press, 1991); Virginia R. Dominguez, *White By Definition: Social Classification in Creole Louisiana* (New Brunswick: Rutgers University Press, 1986); Joel Williamson, *A New People* (New York: Free Press, 1980); L. L. Cavalli-Sforza and W. F. Bodmer, *The Genetics of Human Populations* (San Francisco: W. H. Freeman and Co., 1971), pp. 490-499; T. E. Reed, "Caucasion Genes in American Negroes," *Science 165* (1969), 762-768; P. L. Workman, B. S. Blumberg and A. J. Cooper, "Selection, Gene Migration and Polymorphic Stability in a U. S. White and Negro Population," *American Journal of Human Genetics 15*, 4 (1963), 429-437; Bentley Glass and C. C. Li, "The Dynamics of Racial Admixture - An Analysis of the American Negro," *American Journal of Human Genetics 5* (1953), 1-20; and in general, *Genetic Abstracts* from about 1950. For these references and discussion on this matter I am indebted to Professor Monro S. Edmonson of Tulane University's Department of Anthropology. The PBS three-part television series *Race: The Power of an Illusion* (2003) provides an excellent summary.

[6]This may contribute to an explanation of the phenomenon, noted by Schuman, Steeh, and Bobo (*Racial Attitudes in America: Trends and Interpretations* (Cambridge, Mass.: Harvard University Press, 1985), that in the last twenty years, white support for the principles of equality and fairness for blacks have increased, concurrently with white opposition to the implementation of those principles.

[7] Here the joke characterizing the difference between first-order racism in the American South and North is relevant: In the South, it is said, whites don't mind how close a black person gets, as long as he doesn't get too big; whereas in the North, whites don't mind how big a black person gets, as long as he doesn't get too close. Only the higher-order political discriminator of either region is compelled to deny the existence of the black person altogether.

[8]N. J. H. Dent, *The Moral Psychology of the Virtues* (Cambridge: Cambridge University Press, 1984), 155-160.

[9]This idea of theoretical rationality and theory-building as an innate disposition is given some support by Robin Horton's cross-cultural work. See his "African Traditional Thought and Western Science," in *Rationality*, Ed. Bryan Wilson (Evanston, Ill.: Harper and Row, 1970), 131-171. As I understand Horton's conclusions, the main difference between Western scientific theories and the cosmologies of traditional societies is that the latter lack the concept of modality, i.e. recognition of the conceptual possibility that the favored and deeply entrenched explanation may not be the right one or the best one. They therefore lack the attitude of epistemic uncertainty that leads in the West to the joint problems of scepticism and solipsism. To this extent the stance of intellectual dissatisfaction I am attributing to Kant's epistemology may be culturally specific.

[10]Thus xenophilia in the sense I am defining it should be distinguished from a superficially similar, but in fact deeply perverse form of xenophobia, in which the xenophobe reinforces her honorific, stereotypical self-conception by treating the other as an exotic object of research, whom (like a rare species of insect) it is permissible to examine and dissect from a superior vantage-point of inviolate disingenuity. By contrast, the xenophile acknowledges the disruption and threat to the integrity of the self caused by the other's difference, and seeks understanding of the other as a way of understanding and transcending the limitations of her own self-conception.

[11]By "the Western tradition" in art, I understand not only the Euroethnic canon itself, but also the contributions of colonized, marginalized, or non-Western cultures to it (as, for example, Tahitian art influenced Gauguin, Japanese art influenced Van Gogh, African art influenced Picasso, or American Jazz influenced Stuart Davis).

[12] Thus psychologists sometimes describe intimacy as the case in which you want to either sleep with someone or kill them. Perhaps it is rather that *first* you want to sleep with them, *then* you want to kill them.

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 XLIII* (1984)

_____, "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);

_____, "Durability and Monopoly," *Journal of Law and Economics 15* (1972)

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)

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* (1955)

_____, 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), 500-33

_____, *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* (1953)

_____, "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)

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 (1954-58)

_____, "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), page ??

_____, "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* (1974), 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' 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*, vol. 50 (1972)

_____,"Radical Interpretation," *Synthese 23* (1974): 331-44. Reprinted in *Philosophical Papers, Volume I* (New York: Oxford University Press, 1983), pp. 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)

_____, *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)

McCloskey, H. M., "A Note on Utilitarian Punishment," *Mind 72* (1963), p. 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* (???: Ashworth, 2005)

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 62* (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)

_____, "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* (October 1975)

_____, "The Independence of Moral Theory," *Proceedings of the American Philosophical Association 1975* (Presidential Address)

_____, *The Dewey Lectures 1980*, *The Journal of Philosophy LXXVII* (1980)

_____, "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 (1972)

_____, *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)

_____ 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)

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)

_____ 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 (1982)

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)