OCCURRENT CONTRACTARIANISM:

A Preference-Based Ethical Theory

by

Malcolm Murray

A thesis
presented to the University of Waterloo
in fulfilment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Philosophy

Waterloo, Ontario, Canada, 1995

DECLARATION

I hereby declare that I am the sole author of this thesis.

I authorize the University of Waterloo to lend this thesis to other institutions or individuals for the purpose of scholarly research.

I further authorize the University of Waterloo to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

BORROWER'S PAGE

The University of Waterloo requires the signatures of all persons using or photocopying this thesis. Please sign below, and give address and date. Thank you.

OCCURRENT CONTRACTARIANISM:

A Preference-Based Ethical Theory

ABSTRACT

There is a problem within contractarian ethics that I wish to resolve. It concerns individual preferences. Contractarianism holds that morality, properly conceived, can satisfy individual preferences and interests better than amorality or immorality. What is unclear, however, is whether these preferences are those individuals *actually* hold or those that they *should* hold. The goal of my thesis is to investigate this question. I introduce a version of contractarian ethics that relies on individual preferences in a manner more stringent than has been in the literature to date. "Occurrent contractarianism," as I have called it, is rooted in our social-psychological state. Given the characteristics we have, and given the social situation in which we are embedded, the best resolve we have of furthering our individually defined preferences is to adopt and adhere to a moral system. Occurrent contractarianism remains true to the original contractarian insight; that morality is a rational institution, capable of being designed for and adhered to even by non-tuistic rational beings following merely their own occurrent preferences.

ACKNOWLEDGEMENTS

Above all else, I am indebted to Jan Narveson. He has helpred me in many ways and my gratitude to him exceeds the boundaries of this formal acknowledgement. I wish also to thank Larry Haworth and Paul Thagard for their unwavering support and encouragement, even in pursuits tangential to their own. Peter Vallentyne and Mike Ross agreed to act on my committee on short notice and I very much appreciate that. Most of what I have gleaned from game theory I owe to Paul Viminitz. Various conversations I have had over the years with Craig Beam, Andre Blom, Brian Fleming, Carl Hahn, Louis Groarke, Alix Nalezinski, Doug Mann, Darryl Pullman, and Jason West have been extremely helpful in shaping the current thesis. Behind the scenes are a number of people whom I also wish to thank. They include Debbie Dietrich and Linda Daniel for their maintaining good humour and friendship while supplying on-call administrative assistance. Dave DeVidi and Morgan Forbes offered timely advice to make the process easier. My gratidude also extends to the Philosophy Graduate Student Association, the University of Waterloo Scholarship fund, and the Social Sciences and Humanities Research Council of Canada for their generous contributions to my academic career.

Closer to home, I must give special thanks to my wife, Pat Murray, and my daughter, Emma Murray, for their enduring economic hardship with unparalleled patience. My aunt, Jean Wild's generous patronage purchased the computer upon which this thesis was written and to her I give my sincere thanks. To my father, John Murray, I owe gratitude for my entire education, and from my mother, Joan Murray, I received the undying encouragement to pursue my dreams.

DEDICATION

I dedicate this thesis to Joan Murray, who died, unfortunately, prior to its coming to fruition.

TABLE OF CONTENTS

INTRODUCTION x

PART I
PRELIMINARIES 1

1. CONTRACTARIANISM: AN OVERVIEW 2
- 1.1 The Nature of Moral Theory 2
- 1.2 Contractarianism 4
- 1.3 Some Criticisms 8
- 1.4 Hypothetical Contractarianism 10
- 1.5 Applications 16

2. THEORY AND ANTI-THEORY IN ETHICS 19
- 2.1 The Scope of Morality 20
  - *2.1.1 Broad and Narrow Ethics* 20
  - *2.1.2 Justice* 23
  - *2.1.3 Ignored Virtues* 29
- 2.2 The Non-Cognitivist Critique 31
  - *2.2.1 Reason and Passion* 31
  - *2.2.2 The Evils of Reductionism and the Obliging Stranger* 33
- 2.3 Self-Defeating 37
  - *2.3.1 Communal Self-Identity* 39
  - *2.3.2 Inevitable Cultural Biases* 43
- 2.4 Summary 45
  - *2.4.1 Where to from Here?* 46

PART II
SELF-INTEREST AND PREFERENCE 47

3. SELF-INTEREST 48
- 3.1 Duty Versus Interest 48
- 3.2 Rational Self-Interest 53
- 3.3 The Attraction of Self-Interest 54
- 3.4 Kant 56
- 3.5 Moral Reasons 61
- 3.6 The Nature of Self-Interest 64

4. CONSIDERED SELF-INTEREST 70
4.1 Frankfurt's Model 71
4.2 Gauthier's Considered Self-Interest 75
4.3 Schmidtz's Reflective Rationality 80
4.4 Taylor's Strong Evaluation 84
4.5 Summary 94

5. OCCURRENT PREFERENCES 96
5.1 Occurrent Preferences 96
5.2 Hedonism and Death 101
5.3 Time Independence 104
5.4 Summary 110

6. INSTRUMENTAL RATIONALITY 111
6.1 The Instrumental Model of Rationality 111
6.2 Information 116
6.3 Intransitivity and Indifference 126

PART III
OCCURRENT CONTRACTARIANISM 133

7. OCCURRENT CONTRACTARIANISM 134
7.1 Occurrent Contractarianism 135
7.2 Hypothetical Contractarianism 138
7.3 Sensitive Standard-Based Contractarianism 142
7.3.1 Drug Rehabilitation 143
7.3.2 Surprise Parties 145
7.3.3 Suicide 146
7.3.4 Unconsciousness and Consent 147
7.4 Further Constraints? 148

8. HOW PREFERENCE-BASED CONTRACTS ARE BINDING 152
8.1 Promises, Debts, and Assurance 154
8.1.1 Defaults 156
8.1.2 Spot and Forward Contracts 157
8.1.3 Assurance 158
8.2 Enforcement 159
8.2.1 Formalism 160
8.2.2 Enforcement 162
8.2.3 Realism and Utilitarianism 162
8.3 Agreeing to Enforcement 166
8.4 A Paradox Resolved 167
8.5 Summary 168

9. SELF-EFFACEMENT 171
9.1 A Game Theoretic Approach 173
9.1.1 Gauthier's Model 174
9.1.2 Danielson's Pluralistic Model 179
9.1.3 Chicken 182
9.1.4 The Threat Game 185
9.2 Results 190

10. COMPLICATIONS FOR GAME THEORY 193
10.1 Transparency 194
10.2 The Making of the Translucency Formula 199
10.3 Outcome Values 206
10.4 Population 208
10.5 King Breakers 215
10.6 Computational Costs 220
10.7 Conclusion 225

PART IV
SOCIAL & POLITICAL IMPLICATIONS 226

11. THE STATE AND PROPERTY 227
11.1 Hobbes 227
11.2 Locke 231
11.3 Limited Sovereignty 235
11.4 Justification for the Limited State 239
11.5 The Schism 245
11.6 Property Rights 247
11.7 Summary 250

12. WHY I AM NOT A LIBERTARIAN 252
12.1 Libertarianism and Contractarianism 253
12.2 Self-Interest Revisited 255
12.3 Paretianism 258
12.4 Occurrent-Based Libertarianism 266
12.5 Standard-Based Libertarianism 269
12.6 Summary 276

BIBLIOGRAPHY 277

INTRODUCTION

There is a problem within contractarian ethics that I wish to resolve. It concerns individual preferences. Contractarianism holds that morality, properly conceived, can satisfy individual preferences and interests better than amorality or immorality. What is unclear, however, is whether these preferences are those individuals *actually* hold or those that they *should* hold. The goal of my thesis is to investigate this question.

In my dissertation, I introduce a version of contractarian ethics that relies on individual preferences in a manner more stringent than has been in the literature to date. One of the fundamental tenets of contractarianism is that morality must be individually motivating. The merit of this approach is that contractarianism is committed to keeping a wary eye on human nature; it is a psychologically realist position. What humans are psychologically capable of matters. The problem is that the concept of morality is often depicted as being

precisely antithetical to individual self interest. Morality is designed partly to *curb* self-interest. It appears to be in my interest to steal, but it is immoral to do so. If rationality is defined as a tool to maximize individual preference, *whatever that preference may be*, then there is little hope, evidently, of linking morality to individual motivation. Recent contractarian answers to this problem focus primarily on the claim that it is not *really* in my or anyone's self-interest to steal (or cheat, lie, murder, etc.). In the long term, given the social context of human lives, one does better, individually defined, to live within and according to a moral structure. David Gauthier (1986) believed that so long as an ethical theory appeals to *considered* self-interest, we have a solution to the paradox of constraining self-interest on self-interested grounds. My criticism of appeals to considered self-interest is that morality becomes *standard-bound*, not *preference-based.* What counts as "considered" depends on standards *independent* of the individual's preferences. If we appeal to standards that determine what interests individuals *should* possess, then it follows that a particular moral rule is really in a person's self-interest to abide (not merely to have others abide) *whether or not that person thinks so*. Appealing to standards that have no necessary connection to individual preferences is to abrogate the *raison d'etre* of contractarian ethics. Consider: a perfectly sensible question is to ask in what way is this standard more motivating to me than any given *actual*

preference I may happen to have. If contractarians wish to remain committed to psychological realism, the answer must be: "None whatsoever." And this is a problem for contractarians. My solution is to appeal to one's *occurrent* interests. Occurrent interests are longer standing than mere whims, but the direction and value of these occurrent interests is to be determined exclusively by the agent herself.

My thesis is divided into four parts. The first concerns two preliminary chapters addressing the issue of what contractarian moral theory is and to which ethical tradition it applies. The meat of the thesis begins in part II. The emphasis in part II is to flesh out what should properly count as self-interest and rationality. Part III introduces what I call "Occurrent Contractarianism." It takes the lessons we have learned about preferences and rationality from part II and shows how we can derive a system of moral strictures and obligations based on occurrent preferences. Part of this endeavour makes use of recent work in game theory. Part IV investigates the political and social implications of occurrent contractarianism, focusing predominantly on its link to libertarian thought.

PART I: PRELIMINARIES

Chapter 1 introduces contractarian moral theory and can be safely skipped by anyone already familiar with contractarian thought. Chapter 2 responds to

modern critiques of the ethical tradition in which contractarianism is part. The focus of my thesis remains solely within the confines of contractarian thought. My main concern is what shape contractarian theory should take. The question whether anyone should accept contractarian theory, whatever its shape, thus remains for the most part ignored. Part I and particularly chapter 2 is my attempt at remedying to some degree this question.

PART II: SELF-INTEREST

Part II explores the complicated issue of self-interest. For contractarians to remain loyal to their roots by appealing to individual self-interest, we cannot have ethicists determining in isolation what any given individual's motivations *really* are. There must be a clear tie to what the individual agent's motivations actually are if there is to be any appeal to them. Preferences must be weighted by the individual herself. Chapter 3 looks at the traditional contractarian account of self-interest. Chapter 4 surveys modern accounts that attempt, or so I argue, to objectify self-interest so that it easier accords with moral theory. Chapter 5 explicates what I mean by occurrent preferences; those preferences upon which I build my contractarian theory in part III. Chapter 6 investigates some problems

with the traditional model of rationality upon which preference-based theories of morality are largely dependent.

PART III: OCCURRENT CONTRACTARIANISM

Part III explores the equally complicated issue of how this preference-based notion of self-interest relates to morality. Reliance on standard-bound preferences can solve prisoner dilemmas more easily than occurrent-based versions. This is so simply because the dominant preference-bound strategy is to defect. If we can appeal to the sorts of preferences we should have, this can easily override our more short-sighted straightforward maximizing strategies. Troubles, however, begin. So long as we leave it to the individual to rank her own preferences there is no guarantee that they will match what an ethicist may have deemed appropriate. It would appear that as much as we want a principle to determine when to appeal to standard-bound preferences and when to appeal to occurrent or actual ones, the more simply incompatible the two sorts of preferences are. One is a preference *simpliciter*, and the other is what a rational preference *should* be. The resolution is found within recent work in game theory. What game theoretic results show is that we can reach an (albeit minimal) ethical domain through occurrent preferences by the added appeal to *agreement*. It is this feature that makes contractarianism the successful theory

compared to other theories that also seek to ground morality in rational self-interest. Occurrent contractarianism is not a form of ethical egoism. It does not say simply: “Do whatever is in your occurrent interests." This is the state of nature, after all, and contractarians are anxious to escape the state of nature. As Gauthier argues, the moral sphere is not a *parametric* game, where actors take themselves to be the sole variable in a fixed environment. Rather, agents interact with others, and interaction involves *strategic* choice. Game theory is the tool to help us investigate the ramifications of strategic choice. What is in my interest must be tempered by necessity with what is in your interest. If we wish to interact, and we assume we both see and want the benefits of mutual cooperation, our individual interests must be compromised somewhat. Occurrent self-interest, then, is already necessarily constrained by contractarian principles since contractarianism emphasizes the furthering of self-interest *within the confines of an interpersonal arena*. "Occurrent contractarianism," as I have called it, is rooted in our social-psychological state. Given the characteristics we have, and given the social situation in which we are embedded, the best resolve we have of furthering our individually defined preferences is to adopt and adhere to a moral system. Contractarianism thus, remains true to its original insight; that morality is a rational institution, capable of being designed for and adhered to by non-tuistic rational beings following merely their own occurrent preferences.

The layout of Part III is as follows: Chapter 7 introduces the theory of occurrent contractarianism. Chapter 8 explores how a moral theory based on occurrent preferences can nevertheless yield duties, obligations, and the legitimate enforcement of those duties and obligations. Chapter 9 presents a summary of recent game-theoretic results that help support the conclusions of chapter 8. The contents of Chapter 10, focusing on the problems and limited applicability of game theoretic results, qualifies the findings raised in chapter 9. Chapter 10 is the most technical of the chapters and can be skipped without loss of continuity.

PART IV: SOCIAL AND POLITICAL IMPLICATIONS

Part IV of my thesis explores the social and political implications of occurrent contractarianism. Of particular interest is my claim that occurrent contractarianism cannot serve as a foundation for libertarianism. This may seem surprising since there is much overlap in policy endorsement between contractarianism and libertarianism. After all, if it is a matter of occurrent interests that decide cooperative strategies, we cannot be forced to provide the satisfaction, or the means to satisfaction, of others' interests that we do not have unless we can be compensated sufficiently. Good consequences are defined by the individuals themselves on contractarian grounds. The procedural focus of

contractarianism means that contractarians will not necessarily adopt whatever political system leads to the best, objectively defined, results. Given this, it is beside the point for libertarians to tell us what those best results are -- at least, if their intent is to show that libertarianism is grounded on contractarianism.

Chapter 11 explores the schism between Locke and Hobbes. Hobbes represents the contractarian in this debate, and Locke the libertarian. This chapter is not intended to be simply an historical account. Hobbes is wrong in a fundamental way and I attempt to fix his political conclusions based on what he could legitimately claim if he were an occurrent contractarian. Both Locke and this altered Hobbes will praise limited government. The focus of the chapter, however, is to show that they will agree for *very different reasons*. These different reasons carry over to the distinction between modern libertarianism and occurrent contractarianism, which is the focus of the final chapter of this thesis, chapter 12.

SHORT-LIST BIBLIOGRAPHY

The following are the major works to which I have paid particular attention.

Danielson, Peter, *Artificial Morality: Virtuous Robots for Virtual Games*, London: Routledge, 1992.

de Jasay, Anthony, *Social Contract, Free Ride: A Study of the Public Goods Problem*, Oxford: Clarendon Press, 1990.

Ewin, R. E., *Virtues and Rights: The Moral Philosophy of Thomas Hobbes*, Boulder: Westview Press, 1991

Gauthier, David, *Morals by Agreement*, Oxford: Clarendon Press, 1986.

Hampton, Jean, *Hobbes and the Social Contract Tradition*, Cambridge: Cambridge University Press, 1986

Hobbes, Thomas, *Leviathan*, Buffalo, New York: Prometheus Books, 1988.

Locke, John, *The Second Treatise on Civil Government*, Buffalo: Prometheus Books, 1986 [Originally published as *Two Treatises of Government,* 1690]

Parfit, Derek, *Reasons And Persons,* Oxford, Clarendon Press, 1987

Schmidtz, David, *Rational Choice and Moral Agency*, unpublished manuscript, 1994.

PART ONE

PRELIMINARIES

Chapter One

CONTRACTARIANISM: AN OVERVIEW

1.1 THE NATURE OF MORAL THEORY

Only recently have we started to build a secular ethic.[1] The prejudice of religion has marred the way for scientific insight ever since its ungainly inception. For a number of us, this prejudice has finally been torn asunder. But what sort of secular ethic are we entitled to get? The question is not, by the way, to ask why we have a moral institution if there were no God. The answer may be that people thought incorrectly that there was a God and that this God would punish people for certain acts. Our concern is: even if morality arose through religious belief, why did the religious elite pick out *these* acts to brandish as right or wrong rather than others. Removing the possibility of divine insight (which would not help matters anyway for we would simply ask why did God pick out these acts rather than others), we assume there must be *pragmatic reasons* for the selection of these acts over others. If the pragmatic reason is simply the furthering the

[1] Derik Parfit, *Reasons and Persons*, Oxford: Clarendon Press, 1987, 453.

interests of the religious elders solely, then this grounding will not do for us now. What we want to know, then, is not simply how ethics came about, but is it a good thing *now* to maintain a moral institution. In other words, what is in it for us to be moral?

It is important to realize that by asking this question I am committed to a purely *reductionist* approach to ethics.[2] I do not care *how* people are moral. Presumably they were taught to be. I want to know in what way morality can be built from purely amoral grounds. Whether or not this is why we are moral, it will help to show the mechanics of morality. The claim I will argue for in this thesis is that it is rational on purely self-interested terms to adopt and to act on moral dispositions.

In this project, I propose to build an ethical theory devoid of *a priori* moral or religious assumptions. I am not unique in this attempt and do not pretend to be. Parfit cited Hume and Nietzsche as two notable atheist ethicists.[3] He did not consider Hobbes, however, because Hobbes of course was a theist. But unlike other theists, Hobbes crafted an ethic that was independent of religious belief.

---

[2] See chapter 2 in this thesis for a defence of the reductionist strategy against the criticisms of anti-theorists.

[3] John Stuart Mill should also be given credit for furthering secular ethics. See, particularly his attack against the intolerance of religious dogmatism in his *On Liberty*, Elizabeth Rapaport (ed.), Hackett Publishing Co. [1859] 1978.

The outcome of that project is a theory we call today "contractarianism." This is the theory that interests me. It interests me more than other secular attempts at morality, for example utilitarianism, because contractarianism is motivating in ways that utilitarianism is not. My project is a refinement of contractarianism. What follows in this preliminary is an overview of contractarian ethics. The remaining part of the thesis will focus on modifications of this skeletal frame focusing predominantly on the nature of preferences and self-interest. I do not pretend that other aspects do not also require further clarification or emendation.

1.2 CONTRACTARIANISM

In brief, Contractarianism is the theory that what is moral is derived through interactions with people. Contractarianism holds that morality serves a purpose, and that is to better enable individuals with disparate aspirations to cohabit peacefully. As individuals prosper within a moral realm, as compared to an amoral realm, it is in each individual's self-interest to live in a moral domain. As the moral domain can not flourish without general obedience to the moral dictates that system advocates, it is, indirectly, in each individual's self-interest to obey the moral dictums. Morality is founded on self-interest.

Since morality is man-made for the purpose of securing the interests of disparate people, what is "moral" must meet individual interests. It is assumed

that what people voluntarily and freely agree to is what is in the individuals' interests. Thus, the contractarian credo may be put thus:

> *In order for an act to be considered moral, all people concerned in the act, including the repercussions of the act if any, must voluntarily agree or can be reasonably assumed they would voluntarily agree.*

Its negative, although redundant, thesis is this:

> *Any act by people that negatively affects others who have not voluntarily agreed to being so affected is necessarily an immoral act.*

Who the concerned parties are is a crucial matter for contractarianism. It is not the case that anyone off the street who whines about an agreement made between people other than herself should be sufficient to render their agreement null or impermissible. The whining must be justified. The dissenter must be affected in the requisite way in order to squelch the agreements of others. The requisite way typically is conceived as harm. The appeal to the harm principle is supposed to rule out busybodies who interfere in others' lives for no good reason. That Barney doesn't like Betty's having a nose ring is not thought to count as warranting the banishing of nose rings. The concept of harm, however, is not clearly defined. Barney may really feel "harmed" by people wearing nose rings, albeit in a non-physical way. It would be simple if contractarians could

speak of only "physical harms" when they speak of "harms," but this is not right. Imposed psychological harms cannot be permissible in moral theory. Some criterion of normalacy is required. If someone is harmed by all sorts of things that normal people are not, this individual must be deemed an unreliable measuring rod of harm. What should count as "normal harm," however, is still admittedly vague.

The intent of contractarianism is to permit a maximum amount of negative liberty. You can do anything you want, so long as your actions either effect no one, or you get the consent (possibly through negotiation and compensation) of those your actions affect. This is an easy principle to apply in countless situations. Contractarian principles determine the procedure to test the morality of any given action. It is important to recognize that contractarianism does not mean that the resulting content of any appropriately applied contractarian procedure is necessarily a universal moral dictum in the way that Kant thought. The condition of universalizability for Kant focused on the content of moral actions. That it is wrong for me to waste my talents means it is wrong for everyone to waste their talents.[4] The universalizability condition for contractarians focuses on the universalizable use of the contractarian

[4] cf. Immanuel Kant, *Grounding for the Metaphysics of Morals*, James W. Ellington (trans.) Indianapolis: Hackett Publishing Co., 1981 [1785], 31 [423].

procedure; not the universalizability of the results of any particular use of the procedure. Whatever meets uncoerced agreement from concerned parties is morally permissible. Acting on what does not meet this condition is immoral. That Sally agrees to have sex with Tom does not mean she agrees to have sex with everyone. Or if Sally decides not to have sex with Tom, it does not (necessarily) mean she regards sex for everyone as immoral. Under the contractarian aegis, if one wonders whether a considered act is moral or not, ask simply, do all parties affected by this act agree, or is it conceivable that they would agree if they were rational. We assume that my taking a stranger's purse is not an act that the stranger would consent to. Theft in general will fail the contractarian test as will everything else that is harmful and non-consensual. Euthanasia should be permitted so long as only those concerned agree.[5] Another example where contractarianism is clear concern cases of homosexuality. So long as all parties involved voluntarily consent to homosexual acts, nothing is immoral about it. To disagree is to explicitly reject contractarian principles. The claim would be that it is an immoral act, whether or not the people involved consent. But claiming

[5] I ignore, although Paul Viminitz urges me not to, the possibility that legal euthanasia will merely increase the likelihood of camouflaged murder. It is a burden, financially if not emotionally, on the descendants to keep grampa alive. Life insurance premiums and wills may also make it profitable for family members to stage unwarranted "mercy killings."

homosexuality to be "immoral" just like that is driven by a mere aesthetic or religious fervour that has no place in contractarian moral theory.[6]

Contractarianism, thus, protects individual liberty; forbids the standard "immoral" acts, such as murder, rape, theft, assault, fraud; and offers fairly clear resolutions to many current social issues. A precautionary note may be necessary. The mere application of contractarian principles cannot resolve all moral debates. Disagreements may be about matters of fact or definition. Whether a fetus is a person that deserves to be granted hypothetical assent to abortion is not a matter the theory of contractarianism can say anything about. Contractarianism is not a panacea to good sense.[7] Typically, moral conflicts involve disagreement on which theory to apply, and not on how to apply it.

---

[6] One might (unsuccessfully, mind you) claim to be a contractarian while disallowing certain homosexual acts on the grounds that consent must be between *rational* or perhaps *normal* individuals. One would then disallow homosexuality on the grounds that homosexuals are abnormal or irrational. To claim it is abnormal and that *therefore* is immoral is simply to beg the question, however. Surely part of morality is to tolerate differences among people, not to exacerbate them. To claim that homosexuality is in some sense "irrational" would also be difficult to back.

[7] This is a point Bernard Williams and Alisdaire MacIntyre do not fully appreciate when they argue that the existence of moral dilemmas proves the implacability of foundationalist normative theories such as contractarianism. See Bernard Williams, *Ethics and the Limits of Philosophy*, London, Fontana Press, 1985, ch. 8 and Alisdaire MacIntyre, *After Virtue: A Study in Moral Philosophy*, Notre Dame, IN: Notre Dame University Press, 1984, 6-18.

1.3 SOME CRITICISMS

Contractarianism is not criticized for what it can do, however, but for what it fails to do. If two people agree to row a boat, and this action does not adversely affect others not party to the agreement, nothing could be immoral about the act, or so contractarians would avow. Many strongly protest. It is often supposed, for example, that one's moral duty is to give positive aid to the suffering. That two people consent to row a boat may then be immoral, according to these objectors, if a third were drowning and they were rowing the boat away from the victim. Morality, it is claimed, must require more than mere agreements; in fact, morality must impose strictures on the content of particular agreements. If so, contractarianism fails to ground morality.

The contractarian motto is that any agreement is legitimate so long as it does not adversely affect others. The question before us is whether or not the drowning victim was harmed by the rowers' failure to rescue him. If the drowning victim had a claim right against these two rowers to be saved, then their failing to save him does adversely affect him. But the victim has no such claim right under contractarian theory, and so cannot be considered to be harmed by the suspect rowers. Within the contractarian tradition, "harm" has been defined in relation to a baseline. If an individual is made worse off than she was, this counts as harm. If you are drowning with your wallet in your pocket, your baseline is the state of

drowning with your wallet. Thus, failing to save you will not count as harming you since you will still be drowning with your wallet. Taking your wallet before you drown will be "harming" you for it worsens your state relative to your baseline.[8] Thus contractarians are consistent in maintaining that rowing away from the victim is morally permissible.

This does not answer the problem, really. The complaint is not that contractarians are inconsistent; it is that they are immoral, or, more formally, that their theory of morality fails to capture morality. Knowing why contractarians claim that the two rowers are not harming the drowning man does not convince many that the two rowers are thereby moral. Some think it is "morally monstrous" that a moral theory will have nothing to say to someone who allows another person to drown.[9]

Defenders of contractarianism are for the most part baffled and embarrassed by this complaint. Their confusion is understandable. The

---

[8] I will ignore the complications that death creates. If you do drown, then your baseline is so low that nothing could worsen your state. Thus, seemingly, taking a dead man's wallet may not be morally inadmissible. If the dead man had a will, however, perhaps that wallet, or its contents, were to be given to someone else, in which case it would be immoral to take the wallet, since it reduces the baseline of the beneficiary.

[9] This is Kai Nielsen's injunction in his "Capitalism, Socialism, and Justice," in T. Regan and D. Van DeVeere (eds.) *And Justice for All*, Totowa, N.J.: Rowman & Allenheld, 1982, 264-286.

complaint leaves the boundaries of discourse in which contractarians speak. They are interested in devising a moral theory devoid of appeals to emotions and intuitions. But this objection throws these intuitions and emotions back into the ring. They refuse to play the contractarians' game. Defenders of contractarianism feel, I suspect, that they are not allowed to have these emotions and intuitions since they are trying to build a moral theory independent of them. But nothing is further from the truth. That we have sympathies to help others will motivate us to help those others. Nothing in contractarian theory will ever forbid acting on one's naturally felt motivations. Simply, they profess that we cannot demand others to act on the motivations that *we* feel. It is an ethic of toleration; not totalitarianism.

1.4 HYPOTHETICAL CONTRACTARIANISM

The currently favoured version of contractarianism may be called *hypothetical contractarianism*.[10] The origin of hypothetical contractarianism stems from the criticisms Hume raised.[11] The grounds of political or moral institutions is certainly

---

[10] Hampton has reservations about Gauthier's hypothetical contractarianism, but she does not venture to explain her reluctance. (Jean Hampton, *Hobbes and the Social Contract Tradition*, Cambridge: Cambridge University Press, 1988, 266, nt.7.) I suspect it is a misplaced concern for his appeal to considered preferences that I too have grave doubts about (see chapter 4 in this thesis).

[11] David Hume, "Of the Original Contract," in Eugene Miller (ed.), *David Hume - Essays: Moral, Political, and Literary*, Indianapolis:

not founded on some *original* contract. Conquest and the wielding of power is a more historically accurate portrayal of the foundation of political obedience. Moreover, even if political sovereignty were founded on an original contract, there is no good reason why we are bound to what would be an ancient and possibly barbaric document.

Different from original contractarianism is *explicit* contractarianism. Whatever people explicitly agree to, so long as there are no negative externalities, is necessarily morally permissible. Anything that harms another that was not explicitly agreed upon is classified as immoral. It is a large concession, really, that Hume recognized explicit agreements are a perfectly legitimate ground for morality. The consent of the people is not only a "just foundation of government...It is surely the best and most sacred of any."[12] The problem is that legitimacy based solely on explicit agreement is not practical on a large scale. We can not expect explicit agreement on every action of the state before we deem an act morally permissible. There are just too many people, and often one needs to act quickly. Political obedience must be justified, if it is to be justified at all, on grounds other than explicit agreement.

LibertyClassics, 1987, pt. II, Essay XII, 465-487.

[12] Hume, 474.

Granting the impracticality of explicit consent, some have understood contractarianism as granting legitimacy to political institutions on the basis of *implicit* consent. Implicit consent is the belief that so long as you benefit from some aspects of a political structure, you have thereby given consent to all other aspects of that political reign. The problem is that this will include something to which, had I known, I would not have given my consent. Implicit contractarianism, at any rate, is similar to explicit contractarianism except it is not dependent on time-consuming verbal or written agreements. Some things we can safely assume. In epistemology, even, it may well be reasonable to assume things otherwise unverifiable. It is a good bet, for example, that my chair continues to exist during those intermissions when no one perceives it. Likewise for certain moral justifications. For example, voting is one means of trying to come close to getting explicit agreement. It is not a particularly successful method. To begin with, my choice is highly restricted at the outset. And secondly, my voting for Smith, on this system, somehow confers my consent to Jones getting elected (when more people, however ill-informed, voted for Jones). Or when I vote for Smith because I like his policy on education, say, society assumes for purposes of convenience that I voted for all of Smith's policies. Once Smith has been voted in, he no longer needs explicit consent for everything that he does, even if it goes against many of my preferences. The

time saved is surely beneficial. Smith can achieve many more things if he only requires my *implicit* consent rather than my explicit consent. Moreover, I can get on with my own pursuits without being interrupted by annoying polls. But the assumptions we make about people's preferences are far less sure than the safe assumptions we make in epistemology. The Laws convinced Socrates of the binding force of implicit agreements in the *Crito*.[13] Because Socrates did not complain when he was born and educated in the state, and did not leave Athens, he thereby ought have no complaint against being executed by the state. This absurd result is enough to cast doubt on grounding morality on implicit agreements.

Being hurt by another's action does not necessarily mean it goes against contractarian principles. If one has agreed to partake in a sport or a business in which being hurt is part of the known risk, that one is hurt will count as an aspect of the agreement. Thus becoming bankrupt because a competitor has out-bid you does not count as being harmed by the competitor. Presumably you have agreed to the framework in which business is transacted, and this includes the opportunity for new businesses to emerge. It is not clear that implicit consent works this way even for business transactions, however. It is questionable

---

[13] Plato, *Crito*, in E. Hamilton and H. Cairns (eds.) *The Collected Dialogues of Plato*, Princeton: Princeton University Press, 1961, 51e.

whether you have agreed to all the risks of, say, working in an industry. Severe, debilitating accidents that could have been foreseen and prevented by one's superiors are not the sort of things to which right-minded individuals would likely consent.

Hypothetical contractarianism solves the problems inherent in the earlier forms of contractarianism. The acid test is in deciding what would be *reasonable* for an individual to have consented to, suitably situated and with sufficient knowledge. It would not be reasonable to accept being blinded in the course of one's work, and thus being blinded can not count as being part of the bargain.

Although Hobbes believed we have a duty to accept everything the sovereign doles out to us, the foundation of the contract is not implicit in this sense. For as Hume points out, the implicit agreement would hold even if some tyrant took over the state "by violence" and people submitted to them "by necessity."[14] There is no philosophical or moral reason for why we are bound to observe our tacit promises. For Hume, allegiance to one's government is not due to an implicit agreement, nor an original agreement. Rather, "allegiance and fidelity are both submitted to by mankind, on account of the apparent interests and necessities of human society."[15] I maintain that this is precisely the emphasis

---

[14] Hume, 475.

[15] Hume, 481.

of contractarianism, notwithstanding the usual utilitarian interpretation of this remark.

The utilitarian flavour is derived in this way: Since it is a good for the aggregate whole, that is its justification. But this does not capture Hume's meaning, given his emphasis on the necessity of the benefits to the individual. "For it is evident that every man loves himself better than any other person."[16] What is required to make this into a utilitarian doctrine is the success of Mill's proof of utility, the attempt to show how from this we ought to pursue the happiness of *everyone.* [17] But this is clearly unsuccessful. The contractarian notion, instead, is that by everyone's pursuing their own happiness, within the confines of allowing others to do the same, this best ensures the happiness of everyone. Gauthier calls this understanding of contractarianism *hypothetical* contractarianism.[18] Accordingly, "systems of property and government are

---

[16] Hume, 480.

[17] John Stuart Mill, *Utilitarianism*, New York: Prometheus Books, 1987, ch. IV, 50.

[18] David Gauthier, *Moral Dealing: Contract, Ethics, and Reason*, Ithaca and London: Cornell University Press, 1990, 53.

legitimized in terms of the consent they *would* receive from *rational* persons in a suitably characterized position of free choice."[19]

Gauthier's claim is that Hume was right in objecting to explicit, original, and implicit versions of contractarianism, but that he left unharmed hypothetical contractarianism. In fact, Hume endorsed a form of hypothetical contractarianism. We can see this clearly, I believe, by examining Hume's conception of justice. Justice for Hume extends no farther than mutual convenience and mutual advantage. This can be gleaned especially from his six cases where justice is an unnecessary virtue.[20] In these cases, justice becomes useless to *the individual*. It is no longer in the individual's interest to agree to oblige the requirements of justice. And we take this to mean, *should* it be rational to agree, then it is morally permissible.[21]

---

[19] Gauthier, 53.

[20] I take exception to the first case. According to it, there is no need for justice if we have everything in abundance. In children, at any rate, this does not follow. Among a thousand identical teddy bears, the child wants only the one she's already picked out. We have, I think, a natural possessiveness of
things, and if so, we may still require property rights (i.e. justice) even if there is a natural abundance of things. This possessiveness does not occur only in children, either. Consider our relation to spouses. The claim that there are "many fish in the sea" is not comforting should one's neighbour take one's spouse. Simply because there is plenty of sexual prowess in one's spouse to go around, it does not follow we do not mind if it does "go around." So long as there is emotional attachment to things, however irrational that may be, superabundance, alone, is not enough to render justice unnecessary.

[21] Gauthier, *Moral Dealing*, 35.

According to Gauthier, government dictums are legitimate hypothetically so long as they are what rational individuals, suitably positioned, would freely choose.[22] Rawls, too, so far as he is a contractarian at all, advocates hypothetical contractarianism when he speaks of what people *would* choose from a fair choice situation. He bases his ethical and political policies not on actual agreement between people, but on agreement reached from behind a veil of ignorance.[23]

That both Gauthier and Rawls adopt hypothetical models of contractarianism belies a stark difference between their approaches. Gauthier's model of hypothetical contractarianism is severely restricted by what individuals would choose under their current preference scheme. Rawls, on the other hand, adopts a wholly unrealistic hypothetical situation, and thus his version of hypothetical contractarianism is harder to motivate us in the real world. If part of the *raison d'être* of contractarianism is to motivate rational self-interested people to be moral, then Gauthier's realist version is preferable to Rawls's ideal version.

---

[22] Gauthier, 53.

[23] John Rawls, *A Theory of Justice*, Cambridge, Mass: Harvard University Press, 1971, 136.

1.5 APPLICATIONS

I believe with Smart that the sole purpose of philosophy is its instrumental value.[24] Most writers focus on the political implications of their ethical theories, and others apply their theories to social problems like prostitution, euthanasia, and same-sex bills. It is not difficult to apply a theory like contractarianism to settle most of these disputes. The simplicity of its procedural application is far superior, for example, to the complex and potentially nonfalsifiable methods of utilitarianism.[25] Because our disputes are still raging, we can infer that it is not because we do not know how to apply our theories, but that we have not yet settled on which theory to apply.

The original motivation of my interest in the philosophy of ethics is what to do with criminals. That we should lock them up is not at issue. If offenders of our moral theories are not punished, our moral theory is useless. My specific concern is with rehabilitation. Criminals by and large are not unaware that

---

[24] J. J. C. Smart *Philosophy and Scientific Realism*, London: Routledge and Kegan Paul, 1963, 1-15. For the view that philosophy has an intrinsic value that should be honoured for its own sake however idle it is to the world at large, see Bertrand Russell, *The Problems of Philosophy*, New York: Oxford University Press, 1969, 153-161.

[25] Depending on how we add up the utility scores, we can claim that killing an innocent baby is morally justified (the parents' happiness is increased due to the increased freedom to pursue their own good) or unjustified (the disutility to the infant as well as the outraged onlookers outweighs the happiness of the parents).

society in general deems murder, theft, assault, fraud and the like wrong. Telling them that it is wrong, then, does no good. Speaking to them of God does no good for any atheist. For that matter, it may do little good for certain theists, as well. Either the saved are already predetermined from birthright or predetermination, or the God is a forgiving God so long as a certain incantation is uttered at one's death bed, no matter how heinous one's deeds. Neither does threatening imprisonment do any good for those who earn peer-reputation points for incarceration, or for those whose status quo is less than life in prison, or for those who believe themselves smart enough to avoid incarceration the next time. What to do then? A strong test of any ethical theory is not to see if it passes muster with staid academicians, but whether it can be successfully applied with the prisoner or would-be prisoner population. Rehabilitating inmates with a contractarian theory is more than just of academic interest. A relatively successful psychological therapy for inmates is called "Reality Therapy" devised by William Glasser.[26] Its principles come close to the philosophical theory I am espousing here, and its success (relative to the generally low success rate with the offender population) gives me some support that this theory can be successfully implemented in the real world. Reality Therapy, in brief, appeals to

[26] William Glasser, *Reality Therapy: A New Approach to Psychiatry*, New York: Harper and Row, 1965.

the interests the inmate has and challenges the inmate to assess his past attempts at satisfying those interests. The plan is to show that his own interests are not being satisfied by his criminal projects, and thus he should learn to change his methods of satisfying his own goals and interests.

That this therapy approach makes no appeal to transcendental notions, but only self-interest, should not be surprising when we are dealing with a low-morality biased population. That it nevertheless works is very informative. It helps show that self-interest and rationality can be the sole basis of individual moral constraints.

Chapter Two

THEORY AND ANTI-THEORY IN ETHICS

There are two contrasting paradigms in ethics: theory-driven ethics (such as Kantianism, utilitarianism, libertarianism and contractarianism) and anti-theory-driven ethics (including virtue ethics, moral realism, non-cognitivism, and communitarianism). In this chapter, I will consider three of the anti-theorists' complaints against theory-driven ethics, focusing primarily on contractarianism. (i) Contractarianism offers a narrow and distorted view of morality by being unable to generate virtues such as generosity, kindness, and compassion. (ii) Contractarianism is unduly restricted for being tied to stringent, antiquated models of rationality. Moreover, rationality is not the only motive of people, and hence any pretensions of offering a normative theory is hopelessly idealistic. (iii) Contractarianism is self-defeating. Contractarians claim universalizability, but they fail to clearly recognize that their theory is itself culturally biased, and hence non-universalizable. In other words, theory ethics is ultimately grounded in the very extant cultural normative ethical practices it was supposed to be *justifying*.

I disagree. In this chapter I attempt to show why the anti-theorists' complaints are misguided.

2.1 THE SCOPE OF MORALITY

2.1.1 Broad and Narrow Ethics

One of the consequences of the contractarian conception of the origin of morality is that the province of morality is restricted to interpersonal relations solely. Let us call this the *narrow* view. How one acts in relations to others may be a moral matter, whereas it is never a moral matter, on the narrow view, when your actions concern yourself alone. Anti-theorists, by contrast, hold a *broad* view of morality. They claim that all we need do is look at the world and we can see that morality extends also into the intrapersonal domain.[27] Charles Taylor, for example holds that any deliberation that requires what he calls "strong evaluation,"[28] will be a moral matter. Strong evaluation consists in our discriminating right and wrong; good and evil; noble and base; and other sorts of contrastive terms.[29] Accordingly, any desire

---

[27] See, for example, Michael Walzer (*Interpretation and Social Criticism*, Cambridge: Harvard University Press, 1987), Michael Sandel, (*Liberalism and the Limits of Justice*, New York: Cambridge University Press, 1982), Charles Taylor (*Sources of the Self: The Making of Modern Identity*, Cambridge: Harvard University Press, 1989), Larry Haworth (*Autonomy: An Essay in Philosophical Psychology and Ethics*, New Haven: Yale University Press, 1986), and Alisdair MacIntyre (*Whose Justice? Which Rationality?* Notre Dame: Notre Dame University Press, 1987).

[28] Charles Taylor, "What is Human Agency?" in *Human Agency and Language: Philosophical Papers I*, Cambridge: Cambridge University Press, 1985, 16.

[29] Taylor, *Sources of the Self*, 4.

that involves strong evaluation is necessarily a moral decision.[30] Since we distinguish right and wrong actions even in our intrapersonal spheres of life, morality must necessarily be "broadly" construed.

I have no qualms that we do deem some intrapersonal desires worthy of who we are, and others unworthy, and one way to articulate this contrastive feeling is to call one "good," the other "bad." What needs to be argued, however, is why the mere use of these otherwise "moral" words is enough to draw the matter into the moral purview without simply equivocating on terms.

Anti-theorists such as communitarians believe that intrapersonal goals ought also to play a part in morality in a way that is not merely derivative of interpersonal goals. If one continually beats up on people, then an individual goal to control one's aggressive urges would be a moral matter; but only derivatively so. The communitarian claim is stronger than this restriction: individual goals that have no influence on one's interpersonal goals also matter morally.[31] I think this "broad" reading runs

---

[30] Taylor, *Sources of the Self*, 4.

[31] One could, I suppose, believe that solely intrapersonal goals might matter morally so long as one held a theological perspective. One ought to be good, say, even if alone on a desert island. In this case, "Goodness" is not defined by interaction with others, but by some internal code. What is "good" may be adherence to God's word, and God's word will clearly extend to both interpersonal and intrapersonal actions. I find such a claim implausible, but it at least has the merit of making sense of a "very broad" conception of morality.

counter to our normal understanding of the scope of morality. Typically we believe that not all recommendations are moral ones. For example, "Select a medium brush for best results," is a good recommendation, but not obviously a moral recommendation.

By contrast, contractarians claim that the narrowness of ethics follows from the observation that people have disparate views of what the good is. Since conceptions of the good will tend to differ, conflicts over what one ought to do will likely arise in interpersonal relations. Morality, then, is strictly an attempt to resolve these interpersonal conflicts. Where no interpersonal relation is, one's conception of the good is all that matters, and so there is no need for an arbiter ("morality" for contractarians). Once one accepts the basic premises of subjectively defined notions of the good, the contractarian notion of "narrow" ethics would seem to follow. If morality also seeps into the intrapersonal domain, as the broad view advocates, then morality limits one's personal thoughts, ambitions, and goals. Such restrictions require justification that is not forthcoming from the anti-theorists. Nor can they explain the irony in Sartre's remark in *The Age of Reason*: "Everybody always had Boris's good in view. But it varied with each individual."

If morality is narrowly construed then the sorts of moral demands will concern our interactions with others solely. What you do on your own is permissible so long as it does not adversely affect others. If morality also seeps into your intrapersonal

domain, as the broad view advocates, then morality limits one's personal thoughts, ambitions, and goals. There will be such a thing as a "good life" that you ought to aspire to, much in the way that Aristotle, for example, conceived it. The problem is, this view of the good life may not be shared by everyone, and worse, not shared by you. On this point, I agree with Mill:

> All errors which he is likely to commit against advice and warning are far outweighed by the evil of allowing others to constrain him to what they deem his good.[32]

2.1.2 Justice

A criticism raised by anti-theorists against the narrow view is that it reduces morality to justice issues. Let us first admit that many of our important decisions occur outside the interpersonal arena. What career I should pursue or how I should best spend my free time are matters that concern me deeply, and yet are matters about which the narrow view of ethics could not care less (unless externalities exist). They do not care about these intrapersonal matters not because they are insensitive brutes, but because they recognize that people differ as to what counts as intrapersonally important. What matters morally for the narrow view concerns only

---

[32] John Stuart Mill, *On Liberty*, Hackett Publishing Co. [1859] 1978, 75.

that area of "overlapping consensus," to use Rawls's terms.[33] An overlapping consensus is whatever reasonable people can agree on. To claim that matters beyond this consensus ought to be politically enforced is to impose the values of a sub-group on everyone. On the narrow view, no society has a right to demand of its populace more than what satisfies this overlapping consensus. It is not a necessary condition that an overarching consensus is minimal. It may be that everyone in a closed society believes that it is everyone's duty to profess Catholicism at every level of society; in schools and courtrooms for example. The overlapping consensus then would permit the church to overrule the state. Appeals to psychological realism, however, plus the desire to peacefully associate with more than merely one's own cultural group, precludes such thick conceptions of the overlapping consensus. The narrow view claims that the overlapping consensus is minimal and hence moral and political demands will also be minimal. The consensus will entail neutral matters of justice and little else.

Communitarians can accept Rawls's notion of overlapping consensus. Simply, they will deny that this overlapping consensus is so narrow. Why is this overlapping consensus *simply* the justice answer, they will ask. There is *more* than right dealing in our overlapping consensus, they claim. So why say that is all? Why

---

[33] John Rawls, *Political Liberalism*, New York: Columbia University Press, 1993, 10.

not some common *purpose* in this overlapping consensus? Why not a worthwhile goal on which everyone can agree, rather than this mere conflict resolution procedure?

The question is sensible, it appears to me, only if we limit the range of disparate peoples to those who do share certain goals and values. If the overlapping consensus is supposed to encompass all the peoples of the world, a target to which any moral system should aspire, then this rhetorical question is answered simply by noting that the values or the world's people are too distinct and varied to allow overlapping consensus of such breadth. Rawls avoids this problem, admittedly, by appealing to the overlapping consensus of only those religions, philosophies, and morals that are *reasonable*.[34] Reasonable doctrines, in Rawls's understanding, are only those by which all reasonable people "can cooperate with others on terms all can accept."[35] It is questionable how much import Rawls makes of this conception of "reasonableness." That only "reasonable" people will agree to make justice decisions from behind the veil of ignorance is a position he seems committed to make. But if so, his concept of "reasonableness" is stretched beyond

[34] Rawls, 58-66.

[35] Rawls, 50.

plausibility. As blandly put in the text, however, it seems an acceptable criterion; we do not expect agreement from unreasonable people.

Rawls was concerned only with political conceptions; not moral matters. It is conceivable, in fact, that Rawls did not consider an overlapping consensus of morality to be possible, since he viewed moral conceptions as belonging to those comprehensive doctrines that had to be left at the threshold of a liberal political theory.[36] Nothing prevents our finding an overlapping consensus among disparate (reasonable) moral theories in the way Rawls had done for political theories, however. Searching for an overlapping consensus will find a paradigmatical core agreement among the disparate peoples. For example, do not (*prima facie*) kill, steal, rape, lie.

If we leave it at this negative conception, however, we can not claim that there is a moral obligation to be charitable, or to save someone from drowning. For this reason, the moral prescriptions appear to be identical to the narrow political strictures of justice; hence the communitarian complaint as well as Rawls's reluctance to apply the same test to moral theories as he does to political justice issues.

---

[36] Rawls, 59-60.

Egalitarians and communitarians will conceive of the scope of justice to be fairly broad; including duties of benevolence and duties of forbearance. Libertarians will conceive of justice as far more restricted than that, advocating merely negative duties.[37] If justice were the sole content of morality, morality would be fairly broad on the egalitarian account and horribly narrow on the libertarian perspective. Although contractarians indeed conceive of justice as being restricted to negative demands of non-interference with liberty, conceiving their view of morality as horribly "narrow" follows only if they *also* adhere to the claim that morality is exhausted by justice. But they need not do so.

A more sensible view is that justice is a subset of morality. We might better see this if we construe justice as that which is demanded. Other parts of morality are not demands but recommendations. To disobey moral commands (justice) will merit punishment. To disobey a moral recommendation, on the other hand, will merit disapproval at most. Of course, if the matter is serious enough, we will do more than disapprove; we will cut off our social intercourse with them. Where social intercourse has mutual benefits, these dissenters stand to lose out.

---

[37] For an egalitarian defence, see Kai Nielsen, *Equality and Liberty*, Totowa, N.J.: Rowman & Allenheld, 1983. For a libertarian account, see Jan Narveson, *The Libertarian Idea*, Philadelphia: Temple University Press, 1988.

What we have not yet shown is where or how we should draw the boundary between recommendations and demands. To tentatively use Kant's distinction, think of demands as categorical. At least, let us say they involve fewer excuses and mitigating circumstances than do recommendations. To recommend something is hypothetical. It depends pretty clearly on one's position, situation, and abilities. Let us call this "context" for short. My claim then is this: Demands are relatively context *insensitive*; recommendations context sensitive. It is a recommendation, rather than a demand, if the action depends largely on your ability to do it. But if it is a demand to do something, it is expected that you are able to do that thing; it is expected that differences in context will not diminish your duty to perform to given demand.[38]

If this is the case, the sort of things we can demand of you must be pretty slight given the context sensitivity of our traits and dispositions. Moral demands must be such that practically *everyone* in any context, can obey.[39] Most of the typical positive duties that have any political or social ramifications are context

---

[38] There will be exceptions that makes this distinction not fully Kantian. For example, killing someone in self-defence may well excuse you from a charge of murder, or lying under certain circumstances may seem morally praiseworthy. We can think of the categorical-hypothetical imperatives as dimensions along a sliding scale; categorical on one end, and hypothetical on the other.

[39] I say this because I accept the universalizability principle of ethics. And I accept the universalizability principle because I view ethics as having the goal of social harmony. Social harmony can not be achieved if people are not treated equally on certain dimensions.

sensitive. Whereas all negative duties are context insensitive. For example, demands to feed the poor cannot be context insensitive; whereas demands to refrain from taking anyone's food can be context insensitive. Now, if demands of justice must be context insensitive in order to be universalized, then demands of justice will tend to be limited to negative demands; duties of forbearance. Everyone can equally obey duties of forbearance; whatever their sex, race, social status, or physical or mental abilities.

If it is the case that justice concerns those things we can demand of everyone and only negative duties are the types of demands we can make of everyone, it follows that justice will be exhausted by negative duties. The typical complaint against this "narrow" conception is this: Doesn't morality clearly demand we help victims, prevent starvation, assist our neighbours and the like? Any moral theory that fails to include benevolence is surely "morally monstrous," if not "evil"![40] But such a reaction is misplaced since here we are talking about what the confines of *justice* are, not what the confines of morality are. To think (for example) that Nozick's entitlement theory captures justice is not to say we are forbidden from helping our neighbours, and, importantly, nor is it to say morality may not in fact *recommend* such forms of benevolent behaviour. A moral theory that denied this

[40] This is Kai Nielsen's pronouncement in his "Capitalism, Socialism, and Justice," in T. Regan and D. Van DeVeere (eds.) *And Justice for All*, Totowa, N.J.: Rowman & Allenheld, 1982, 264-286.

would fail to meet a realistic constraint on what a moral theory should do for us. All a "narrow" concept of justice says is that we cannot *demand* benevolent actions. It is another thing to recommend it, and we may even have good reason to recommend it, let alone reasons that are good from the point of view of the agent.[41]

So far, the narrow view maintains that morality is not reduced to justice issues as communitarians claim, since there is a distinction between moral demands and moral recommendations. At this point, a perfectly sensible question is this: Well, then, what prevents you from saying morality will also include intrapersonal matters? For consider, you have just admitted that moral recommendations need not be context insensitive, as moral demands do. On what grounds, then, do you stop these moral recommendations from entering into the intrapersonal domain? Although I cannot demand that you be less apathetic, there is nothing stopping me from recommending that you be more caring. True this is not a justice issue, but you have admitted that justice issues do not exhaust the moral domain.

Thus, to maintain a narrow view of morality over a broad view, there must be a different sort of distinction to be made than appeals to justice. The narrow view

---

[41] Here I refer readers to David Gauthier's attempt at defending a self-interested morality in *Morals by Agreement*, Oxford: Oxford University Press, 1986.

holds that neither recommendations nor demands are moral (as opposed to non-moral) if the issue involves no direct interpersonal associations.

2.1.3 Ignored Virtues

Anti-theorists have more to complain about than the restriction to interpersonal relations. Annette Baier, for example, argues that theory-based ethics cannot account for the various virtues such as "kindness," "generosity," "gentleness," and "humility." To artificially truncate morality to exclude these virtues is to offer not only a narrow but also a distorted view of moral practices.[42]

The complaint has force if we antecedently assume that theory must conform to folk-ethics of the current society. Then any theory that cannot adequately accommodate all the varied and conflicting doctrines adumbrated by folk-on-the-street must be thrown out. But since the issue revolves around precisely whether or not theories have force independent of its ties to particular folk-ethics, this antecedent assumption merely begs the question. Nevertheless, there is some weight to Baier's remarks. Any ethic that cannot adequately explain the general assent to benevolence, kindness, generosity and the like is to be seriously questioned.

---

[42] Annette Baier, "Doing Without Moral Theory?" in S. Clarke and E. Simpson (eds.) *Anti-Theory in Ethics and Moral Conservatism*, Albany, NY: State University of New York Press, 1989, 26-48.

Narrow ethics does not of course outlaw such virtues. They recognize their importance, but only as further members of the wide ranging human passions that people will find it in their interest to pursue. Narrow ethics carves out a moral space in which these, and other passions, are allowed to flourish.

This does not resolve the problem, however. The fact that some of us feel altruistic, say, does not yet explain why we *endorse* altruism. We do not merely *permit* kindness, compassion, and generosity; we in fact *recommend* it, and often frown upon its absence. It is phenomenologically inaccurate to say we permit someone to rescue a child drowning. Such action is to be wholeheartedly recommended. How can the narrow view of ethics consistently maintain that we should endorse altruism? It would appear they have no grounds to make this claim. Many theory-driven ethics too often explain our *reasons* for recommending altruism by reducing the altruistic feeling to some non-tuistic rationalization. Would it not be better for *you*, they say, if everyone helped others when in need. And then, of course they need to make the further move to show why *therefore* you should be altruistic too. But barring the muddle they get into by this attempt, this does disservice to the phenomenology of why we act altruistically. I don't think: "Ah ha! Now he or his parents or the spectators or God owe me something." Simply, I have the child's interest at heart.

Contractarians do not fall into this muddle, although they, too, appeal to self-interest in explaining this endorsement. It is the same justification for any assurance problem. I have altruistic feelings, and therefore, in a sense, I cannot help acting on them. What I need assurance on is that you too will have altruistic feelings so that I won't be disadvantaged by mine. Arguing in this line does not reduce our altruistic feelings to non-tuistic ones. It does not reduce the *feelings* you have for someone, simply it defines the *reasons* we have for *recommending* those feelings. Of course, my recommending others to be moral is itself a strong feeling, the phenomenology of which is not captured by talk about assurance problems. True enough, but I cannot justify my affecting others by the mere appeal to the fact that I have a certain feeling. I cannot legitimately claim that you must be kind to me because I feel it's right. Mere feelings, or moral sentiments, are sufficient explanation for the intrapersonal sphere, but as soon as you require or recommend others to share those feelings, you need to do more than pound your chest in your attempt at pointing at those sentiments you have and expect others to share. To legitimately influence others requires justification. Appeals to theory is therefore not inappropriate.

2.2 THE NON-COGNITIVIST CRITIQUE

2.2.1 Reason and Passion

Hume argues against the belief that we are motivated to act according to what is rational. Rather, the passions rule. This observation has been used as an attack on theory-driven ethics given their commitment to deriving morality from purely rational principles. The complaint is undeserved, however. Rationality under the contractarian's notion is held to be simply instrumental; a tool to satisfy one's preferences. The preferences, or passions, then, are the master; and reason its slave by serving only the passion's wishes. This understanding of rationality is seen to be in complete accord with Hume's notions. Citing Hume's observations of the link between rationality and the passions as refuting contractarianism is therefore peculiar.

The mere fact that the passions rule reason does not mean that reason plays no role, or even a central role.[43] A theory or anti-theory advocating that what is right is whatever your passions say is right is simply ethical relativism. The problem with relativism in relation to ethics is that it is just false. It is a false doctrine, not merely because it generates contradictions, but that it is false even according to the folk-

[43] This, too, is consistent with Hume: "One principle foundation of moral praise being supposed to lie in the usefulness of any quality or action, it is evident that *reason* must enter for a considerable share in all decisions of this kind; since nothing but that faculty can instruct us in the tendency of qualities and actions, and point out their beneficial consequences for society and to their possessor," (David Hume, "An Enquiry Concerning the Principles of Morals," Appendix I: "Concerning Moral Sentiment," in L.A. Selby-Bigge and P.H. Nidditch (eds.), *Enquiries Concerning Human Understanding and Concerning the Principles of Morals*, 3rd. edn. Oxford: Clarendon Press, 1989, 285).

ethics that the non-cognitivists appeal to. When Albert says, "abortion is permissible under certain conditions," he does not mean that since Jane thinks it is always wrong, it is always wrong. Quite the opposite. He thinks Jane is wrong.

The anti-theorists need to explain how moral sympathy or folk-ethics alone is going to do the job of solving interpersonal conflicts. *Which* sympathies do we call "moral," and why? But the non-cognitivists, not only are unable to answer this question, are not even allowed to ask it. They can do nothing more than shrug or simply shout louder with Mooreian emphasis: "*This* is moral. *This* is not."

But if the appeal is merely to *your* intuitions as to what is moral, why should anyone else accept that -- especially when *their* views differ so much? Resolving this, and justifying coercion requires appeals beyond mere subjective belief. The appeal to rationalism is understood by contractarians as a minimal requirement. What is rational is what will satisfy your preferences given the circumstances in which you are embedded. That some of those circumstances are social ones, it is incumbent upon you to consider the desires of the other agents, desires that may be counter to your own. Ignoring this is done so at your own cost, given the free and rational actions of others. Hence, some degree of compromise, i.e., moral resolutions, is rational. That moral choice is rational means simply it will best satisfy everyone's preferences, *whatever they are*, given the social, historical, and psychological factors in which we are embedded.

To criticize *this* notion of rationality is to criticize the desire of people to satisfy their preferences. And that goes against psychological realism and appeals to folk-ethics.[44]

2.2.2 The Evils of Reductionism and the Obliging Stranger

Many theory-driven ethics, and contractarianism is one, are reductionistic. These attempt to explain why we term some acts moral and others immoral without appeal merely to the emotive force of moral judgements. Ever since Ayer and Stevenson,[45] a growing number of people have challenged reductionist approaches to ethics.[46] Gass put it succinctly: "[P]rinciples ... frequently get in the way of good sense."[47] Gass does not appeal to brute intuition of the Mooreian brand. Rather, he wishes

---

[44] For an admirable account of why ethics should not violate the principles of psychological realism, see Owen Flanagan, *The Varieties of Moral Personality*, Cambridge Mass: Harvard University Press, 1991.

[45] A. J. Ayer, *Language, Truth, and Logic*, New York: dover Publications, [1946] 1952, 102-120. Charles L. Stevenson, "The Emotive Meaning of Ethical Terms," *Mind*, 46, 1937.

[46] Alisdaire MacIntyre, *After Virtue*, 2nd edition, Notre Dame, In: University of Notre Dame Press, 1984. Michael Sandel, *Liberalism and the Limits of Justice*, New York: Cambridge University Press, 1982. Charles Taylor, *Human Agency and Language*, Cambridge: Cambridge University Press, 1985. Simon Blackburn, "Moral Realism," in J. Casey (ed.) *Morality and Moral Reasoning*, London: Methuen, 1971.

[47] William Gass, "The Case of the Obliging Stranger," *The Philosophical Review*, 66, 1957, 203.

to draw our attention to the bare moral transparency of some acts. It is wrong to burn someone alive for no reason. To ask, yes, but what is *wrong* with burning a man alive is to clearly suffer "from a sort of *folie de doute morale*."[48] Any reductionist theory that attempts to explain just what it is about the act that makes it wrong will be more inscrutable than the event it attempts to explain. It will sound simply ludicrous, queerly unfamiliar, out of place, and as Gass avows, merely laughable.

This challenges reductionist theories in general. Rather than starting with principles and developing a moral or political structure from them, Gass and company argue for a non-reductionist approach that begins with clear cases of morality and immorality and to treat these as data.[49] The reductionist approach, however, does much the same. Simply, from these clear cases, an underlying principle is deduced. The merit of the reductionist approach, then, is to find an underlying principle that explains why we treat certain cases as moral and others as immoral. And knowing this will enable a normative theory to be developed. Philosophy, after all, is not a purely descriptive enterprise. The non-reductionist approach is flawed, philosophically speaking, for being confined to a purely

---

[48] Gass, 197.

[49] I should note that Rawls attempted a happy medium between these extremes that might interest those convinced by the Gass-MacIntyre lines of attack. (See John Rawls, *A Theory of Justice*, Cambridge, Massachusetts: The Belknap Press of Harvard University Press, 1971, 48-51.)

descriptive role. If our interest is in normative ethics, the non-reductionist approach is idle.

Defending reductionist approaches in general does not guarantee that the contractarian reduction is successful. If we find cases where contractarians say, "Moral," while our intuitions say, "Immoral," we have reason to doubt the reductions of contractarianism. William Gass offers such a case.[50] His "clear case of immorality," is troubling for contractarians not merely because of their inability to adequately explain *why* it is wrong; but that they cannot even claim that it *is* wrong! Let us see why.

> Imagine I approach a stranger on the street and say to him, "If you please, sir, I desire to perform an experiment with your aid." The stranger is obliging, and I lead him away. In a dark place conveniently by, I strike his head with the broad of an axe and cart him home. I place him, buttered and trussed, in an ample electric oven. The thermostat reads 450^0 F. Thereupon I go off to play poker with friends and forget all about the obliging stranger in the stove. When I return, I realize I have overbaked my specimen, and the experiment, alas, is ruined.[51]

---

[50] The original article, written before Gauthier brought contractarianism back in vogue, actually targeted utilitarianism and Kantianism, predominantly (cf. Gass, 198), or reductionist theses in general.

[51] Gass, 193.

He continues, "Any ethic that does not roundly condemn my action is vicious."[52] Now, since the stranger consented to participate in the experiment, and anything that meets mutual consent is morally permissible according to contractarians, it would appear that contractarianism is therefore a vicious ethic, if it is an ethic at all.

One could complain that the obliging stranger did not, presumably, consent to being hit on the head and burnt in an oven. He consented to partake in some experiment, and most of us who consent to fill out questionnaires in malls or help out in psychology experiments assume that we will not be adversely affected in the process. So long as contractarians speak of "informed consent" in grounding the moral permissibility of an act, contractarianism would not be shown to be vicious after all, since the onus was on Gass to explain to the stranger what the experiment constituted.

Let us side-step this response by altering Gass's case. What if Mr. Gourmand approaches a stranger and asks permission to chop her up and roast her and she, knowing full well the implication of the request, obliges?[53] According to contractarianism this would evidently be morally permissible, but to normal

---

[52] Gass, 193.

[53] I owe this rendition to my colleague, Louis Groarke.

people such an act is morally atrocious, whether it was agreed to or not. So much the worse for contractarianism. Why would anyone defend such a theory?

The answer is this. The cost of making this case transparently clear is its removal from the rather more muddy reality. I admit that under such a scenario, contractarians have no moral right to intercede. Roasting the obliging stranger is a morally permissible event. What it is not, however, is an instance that can properly capture our moral intuitions. The moral outrage we feel is based on our inability to grasp one of the requisite premises of the thought experiment. The scenario demands that we assume the stranger obliges. This is what we cannot fathom. Our normal intuitions are such that the stranger *would not*, if she were at all normal, acquiesce to the horrible request. Being unable to dismiss this normal intuition, we also feel the moral outrage in the continuance of the act. But so long as she does not oblige, or is not in a fit frame of mind to appreciate the request, our moral outrage is justified under contractarianism.

2.3 SELF-DEFEATING

The last of the anti-theorists's complaints that I shall look at stems from the fact that "good" is a varied notion, depending on one's culture, personal identity, and other social, historical, and psychological factors.[54] This is interesting, since this

[54] See for example, Charles Taylor, "The Diversity of Goods," in A. Sen and B. Williams (eds.) *Utilitarianism and Beyond*, Cambridge:

was the insight of Hobbes. Contractarianism is in fact built upon this precise notion. So it is not obvious why it is used as a criticism of contractarianism. The explanation given by the anti-theorists is that "utility," say, or "liberty," or "negative liberty," or "justice," or any of these sorts of abstract concepts represent the contractarian's notion of an absolute, inviolable "Good." Holding absolute universal concepts of the good violates contractarian's own conditions of the subjectivity of "good." Hence, contractarianism is self-defeating.

This response is right in so far as we view these abstract concepts appealed to by contractarians as a substantive value on equal footing with the countless other substantive values that we may share. Unfortunately for the anti-theorists, this is the wrong way of looking at it. It is true that justice is not the only important thing to humans. But this is not the contractarian's claim. It is not even claimed that a particular theory of justice is the only thing disparate groups share in common. Rather, justice is the only thing that best ensures the diversity of things that are important to people. In Rawls's terms, justice is one of the primary goods that enables everyone to have whatever it is they want.[55] A primary or what Paul Viminitz calls an "enabling" good is that which secures the space for subjective goods to be

---

Cambridge University Press, 1982.

[55] Rawls, *Theory of Justice* 72-77; 178-190.

attained. Likewise, liberty is not the only important thing for people, since we can imagine cases of giving up our liberty in order to benefit in other ways. This happens all the time, when we accept jobs, for example. But this does nothing to dampen the contractarian's point, since to give up liberty to take a job, say, is something we should have the liberty to do as well as not to do if such is our desire. If we had no liberty to pursue our passions we would not be able to serve our passions.

Both Taylor's and MacIntyre's main complaint is that contractarians do precisely what they criticize others for doing: foisting their pet values on everyone else. We can now see that they misinterpret the contractarian's concern for primary or enabling goods. The "rule of right" or the "rule of toleration" is not the only important thing to anyone, let alone everyone. The important things are what the rule of toleration protects, not the rule of toleration itself. Justice is not valued for its intrinsic worth. Rather, justice is valued merely as a way of preserving the disparate goods that individuals do intrinsically value.

Some anti-theorists have even claimed that theory ethicists conceive "utility" as a universal good; a good to which all should aspire, a good by which we can evaluate all acts.[56] But this is mistaken. "Utility" is a measurement of individual

---

[56] Stanley Clarke and Evan Simpson, *Anti-Theory in Ethics and Moral Conservatism*, New York: State University of New York Press, 1989, 3.

values and is not itself a value. To think so is to commit a category mistake; similar to assuming that because individual books are written by authors and a library is a collection of books, that therefore the library is written by an author.

2.3.1 Communal Self-Identity

There is another side to the argument that theory-based ethics are self-defeating. One of the maxims for anti-theorists is this: No individual conceives her notion of the good in isolation from other individuals. "One is a self only among other selves."[57] The narrow view, accordingly, demands that individuals are conceived of as atomistic beings utterly independent of one another. And this individualism is an illusion.[58] Individuals have a subjective sense of the good, but do so not as monads, but as members of a community. Thus, conceptions of the good are not individual matters alone, but are part of the collective community consciousness. Conceptions of the good are thus claimed to be communal matters.[59]

---

[57] Taylor, *Sources*, 35.

[58] Taylor, *Sources*, 39.

[59] My complaint is with another aspect of this argument, but it is also worth noting that this connection is invalid. Simply because conceptions of the good are part of the collective community consciousness, it does not follow that conceptions of the good are communal matters. That is to say, so long as we take "communal matters" to imply that the community can decide moral issues by vote. But such design on morality is contradictory to the origins of sanctified moral issues through collective community consciousness.

Taylor is not the originator of this view: he is merely one of a long line of communitarian thinkers, including Rousseau and Bradley.[60] But all such communitarian proposals succumb to the same criticism; namely, the mere fact that we are influenced by the community in which we live has little to say about how we *ought* to act.

I cannot deny that we are influenced by our community. If our strong evaluations are dependent on what sort of personal identity we have, then it follows that our strong evaluations will be influenced by these external social factors. But this argument can have effect on the narrow view only if we exaggerate the claim beyond plausibility.

For the social identity argument to have any force against the contractarian position, the argument would need to take either of the following two forms: (i) We are constituted by our community space. Thus, what really matters to me will be to a large extent the result of my communal space rather than my particular individuality. But if my identity and strong evaluations are constituted by my communal space, then it follows that others in my communal space will also share my intrapersonal values and goals. If so, the overlapping consensus will be much

---

[60] Jean Jauques Rousseau, *The Social Contract*, Maurice Cranston (trans.) New York: Penguin Books, 1982 edition [1762]. F. H. Bradley, "My Station and its Duties" in *Ethical Studies*, London, 1876.

broader than what liberals conceive. The overlapping consensus will also encompass our strong evaluations.

This line of reasoning is amiss. We need not deny that social influences greatly affect our strong evaluations to deny that we should expect any homogeneity from this fact. The reason is our identity is constituted by a multiplicity of social variables. That we are socially influenced, or even socially constituted, is not to say we are influenced or constituted by only one thing. The term "social" is a catch-all phrase for a multiplicity of factors, including our social positions, our career choices, upbringing, surroundings, associations, etc. Although there is some plausibility that any one factor will have an influence on many people, it goes against probability that many individuals will be equally affected by all the same variables together. So although we are socially affected (or stronger: constituted) it does not follow that we will share the same values. We should not expect homogeneity of values among individual social beings simply on the basis of common social factors.

The second form of the communitarian argument is this: (ii) We are constituted by our community space. Thus, what really matters to me will be to a large extent the result of my communal space rather than my particular individuality. But since it matters to me what my identity is, it matters to me what my communal space is. Therefore, we should not simply allow *laissez-faire* mentality to rule our

political conceptions. We should instead design and implement a certain *sort* of society; one that is conducive to the personal identities we desire.

This argument advocates strong positive rights and its antecedent obligations. It argues against contractarian notions of finding a happy medium among competing conceptions of the good, and argues instead that some conceptions of the good should reign supreme. The problem here is conceptual. Again, we need not deny that social influences greatly affect our personal identity, but this line of thinking goes too far. If my identity is that which my community creates, and my identity is that with which I do identify, then it is difficult to imagine ever criticizing the community in which I live -- particularly on the grounds of its effect on my identity. Conceptually, there could be no identity crisis if identity is constituted by one's community, and we are all immersed in some community. On what grounds can we possibly desire revisionist change to our present communal structure? It can not be on the grounds of our *current* personal identities, if what we mean by identity is that which we *do* in fact identify with. To conceive of a better society, or a better identity, can only be imaginable if the present community has less influence on our identities than this strong argument claims.

Either interpretation, then, (i) or (ii), of the communitarian identity argument is unsuccessful when laid out in the above manner.

2.3.2 Inevitable Cultural Biases

A similar argument to the communal self-identity rejoinder is that contractarian doctrines have an inevitable bias that destroys their liberal aspirations. Rawls, for example, claims the goal of liberal politics is to achieve toleration among disparate cultural groups.[61] To achieve a proper overlap of the varying (reasonable) comprehensive doctrines of the different cultural groups represented within a given society, the political structure cannot pander to any one cultural group. It must be culturally neutral. Such an ideal, anti-theorists complain, is impossible. Even aspiring to a cultural neutrality is itself part of the individualistic culture to which Rawls and other contractarians belong. In this case, they are merely imposing their cultural norms on everyone else, and brandishing them as the only *reasonable* ones.[62] They cannot help but violate their own principles.

Lessons we have learned from the natural language school should put this complaint to rest.[63] To alter a term so that its negation is an impossibility is to have altered the term for the worse. Bouwsma noted the puerility of claiming that

---

[61] Rawls, *Political Liberalism*, 4. It is more than this, but this will do.

[62] Although Taylor, Haworth, and Sandel make this claim, see also, more recently, Margaret Moore, "Political Liberalism and Cultural Diversity." Unpublished manuscript, 1995.

[63] See for example, John Austin, *Sense and Sensibilia*, G. J. Warnock (ed.) Oxford: Oxford university Press, 1962.

everything is an illusion. Whether everything is an illusion or not does not alter the fact that we make a distinction between reality and fiction, dreaming and being awake, or the distinction between real barns and papier-maché barns.[64] Dennett, in a similar vein, argued against the philosophical concept of determinism that precludes the distinction of being chained to one's chair and not being so bound.[65] We can make a similar observation of the communitarian's claim that everything is essentially culturally bound. Perhaps so, contractarians can respond, but who cares. It does not deny the distinction contractarians make: that political or moral institutions should try as far as possible to accommodate everyone, no matter their particular cultural background.

How best to do this? The liberal answer is to adopt negative rules. This, they argue, best permits the pursuit of whatever one's disparate goals are, with the sole restriction of not interceding in others' pursuits. That a fundamental theist believes his pursuits are thwarted because the crucifix is not hung in public schools or court houses is to be deemed unreasonable not because it doesn't accord with the liberal's secular culture, but because it doesn't accord with *any* other culture.

---

[64] O. K. Bouwsma, "Descartes's Evil Genius," in O. K. Bouwsma, *Philosophical Essays*, Lincoln, Nebraska: Nebraska University Press, 1965, 85-97.

[65] Daniel Dennett, "I Could Not Have Done Otherwise: So What?" *Journal of Philosophy*, 81, 10, 1984, 553-565.

The rejoinder that leaving public schools and court houses bereft of religious symbolism accords only with the liberal's sense of secularism is to misconstrue the absence of cultural symbols for being itself a cultural symbol. The removal of cultural symbols from political offices is best viewed as a fair resolve given the existence of competing cultures. It is the anti-theorist's complaint against this resolve that reveals bias.

2.4 SUMMARY

Anti-theorists have argued against theory-driven ethics for the following reasons. (i) Theory ethics offer a narrow and distorted view of morality by being unable to generate virtues such as generosity, kindness, and compassion. To this I have argued that contractarians have not excluded these virtues from the moral purview; merely their allowance and recommendation is derivative of the main purpose of morality: to allow peaceful coexistence notwithstanding disparate views of the good.

(ii) Contractarianism is unduly restricted for being tied to rationality, and that rationality is not the only motive of people. Hence any pretensions of offering a normative theory is hopelessly idealistic. This argument is misinformed about the role of rationality in ethics within contractarian theory. The contractarian notion of rationality is consistent with Hume's, not inconsistent with it.

(iii) Contractarianism is self-defeating. There are a number of variations of this argument, but the general trend is to claim that the theory is itself unconsciously motivated by the particular culture in which the theoreticians are embedded. To the extent that *everything* is biased by culture to some extent, this complaint is idle. To the extent that we can nevertheless distinguish enabling goods from raw goods, this complaint is false. The hope of an overlapping consensus of disparate groups is not idealistic if what we expect to find are negative principles of toleration rather than some robust notion of good. They are different types of things and confusing them is what has caused the recent epidemic of anti-theorists.

2.4.1 Where to from Here?

To broaden morality into one's private affairs is to broaden societal strictures into one's private affairs. The liberal tradition narrows the scope of morality with the intention of thus allowing individuals to pursue their own disparate goals with the minimum of external interference. Problems, however, remain for narrow ethics. A key problem is how morality is conceived to be in the interest of the individual. We have objected against the broad conception of morality on the grounds that broad ethics cannot necessarily appeal to the self-interested nature of disparate individuals. But how can narrow ethics do any better?

As stated, this question is too broad. There is a wide range of advocates of narrow ethics and depending on the ethical theory we will have differing answers. Contractarianism is one of the liberal theses that views morality narrowly. What I am concerned with is how contractarians ought to answer this question. How can contractarians successfully derive a morality from the self-interested nature of disparate individuals? Parts II and III of my thesis address this question.

PART TWO

SELF-INTEREST AND PREFERENCE

Chapter 3

SELF-INTEREST

If morality is to be at all useful it must cater to human self-interest. An ethical theory that fails to cater to self-interest, on the contractarian view, fails to be psychologically realistic. This claim can be criticized on two levels. (i) Is it the case that morality can hope to pander to self-interest? And (ii) Is it the case that humans are motivated solely by self-interest? In Part II, I wish to examine the second question. In short, my answer is: Yes, so long as we understand what we mean by "self-interest." In Part III, I will investigate the first.

3.1 DUTY VERSUS INTEREST

Many have interpreted the contractarian tradition as claiming that the sole moral principle, the principle upon which all other moral dictates are based, is the obligation to keep one's agreement. Socrates tells Crito that his execution is a moral obligation since he had, in effect, agreed to it (although vicariously through an implicit agreement).[66] Such a notion can also be seen in Hobbes's Third law of

[66] Plato, *Crito*, in E. Hamilton and H. Cairns (eds.) *The Collected Dialogues of Plato*, Princeton: Princeton University Press, 1961, 51, e.

nature: "*That men performe their Covenant made*,"[67] as well as such statements as: "and that he *ought*, and it is his Duty, not to make voyd that voluntary act of his own: and that such hindrance is Injustice, and Injury...."[68] It is no wonder, then, that A. E. Taylor interprets the contractarian as claiming:

> ...it is always to my interest to conform to the law. And to say that this is to my interest is equivalent to saying that it is my duty; my duty, in fact, means my personal interest, calmly understood.[69]

I believe this interpretation corrupts the whole enterprise of basing ethics on self-interest. I have no "duty" to follow my interest. Simply, I have an interest in doing so. In the State of Nature no duty exists. Yet interests exist. Consequently the two can not be equated in the way Taylor prescribes. Should they be, then since it is the case that we have an interest in electing a sovereign, we are bound to do so. But this is not Hobbes's position. Rather, we are bound to obey the sovereign because we have *contracted* to obey him.

---

[67] Thomas Hobbes, *Leviathan*, Buffalo, New York: Prometheus Books, 1988, xv (74).

[68] Hobbes, *Leviathan*, xiv (68).

[69] A.E. Taylor, "The Ethical Doctrines of Hobbes," *Philosophy*, XIII, 1938, 406. [Reprinted in Bernard Baumrin (ed.), *Hobbes's Leviathan: Interpretation and Criticism*, Belmont, Cal: Wadsworth Publishing Co., 1969, 35-36.]

Taylor is wrong, then, but so too is Hobbes here. If it is the case that the sovereign demands what is not in one's self-interest, largely understood, then there is no duty to obey him. Otherwise, as Hume noted, there is a difficulty for those who wish to base contractarian ethics on self-interest. "We are bound to obey our sovereign, it is said; because we have given a tacit promise to that purpose. But why are we bound to observe our promise?"[70] On Taylor's interpretation, the contractarian can only claim "It is our duty." It is for this reason that Taylor saw in Hobbes's ethical doctrine "a very strict deontology, curiously suggestive ... of some of the characteristic theses of Kant."[71]

The "duty to stick to one's agreements," as explained in the natural laws of Hobbes, is the wrong interpretation of contractarianism. The reliance on self-interest that is crucial to contractarianism is instead best expressed by Hume: Only when everyone may reasonably expect to benefit does justice or obedience obtain.[72]

---

[70] David Hume, "Of the Original contract," in *David Hume: Essays: Moral, Political, and Literary*, Eugene Miller (ed.), Indianapolis: Liberty classics, 1987, 481.

[71] A.E. Taylor, 36-37.

[72] Hume, 480-481. Gauthier recognizes the irony here, since Hume is avowedly a critic of contractarian thought (David Gauthier, "David Hume," in David Gauthier, *Moral Dealing: Contract, Ethics, and Reason*, Ithaca and London: Cornell University Press, 1990 45-76.)

Contrary to Hume's explicit avowals, Hume's rowboat analogy is a classic example of contractarian thinking. In the rowboat analogy, two people cooperate for mutual benefit. Hume's point in raising the example was to show that no explicit contract was ever made.[73] The important concession to contractarian thought, however, is in Hume's arguments that the moral convention determining each rower's action is based on individual self-interest. Each rower does better if they both conform to the convention of shared rowing. What Hume failed to see is the fact that no contracts being made should not bother contractarians, although it does argue against the need for a sovereign to overlook such cooperative tasks. But contractarians do not require a sovereign, contra Hobbes.

Conventions exist precisely because they work. True, there may be a time when they no longer serve their original purpose, due to further information or changed circumstances, yet they linger anyway. Religion, for example, comes to mind. But generally, conventions come into being for contributing in some way to the security of individuals. It is the legitimacy of these conventions that confer

---

[73] David Hume, "An Enquiry Concerning the Principles of Morals," In L.A. Selby-Bigge and P.H. Nidditch (eds.), *Enquiries Concerning Human Understanding and Concerning the Principles of Morals*, 3rd. edn. Oxford: Clarendon Press, 1989, app. III. And also David Hume, *A Treatise of Human Nature*, L.A. Selby-Bigge (ed.) 2nd edition revised by P.H. Nidditch, Oxford: Oxford University Press, 1978 [1888], bk III, pt. II., sec. III.

consent as Hume points out,[74] rather than the reverse. Consistent with empirical observation, people would not consent to that which is not in their self-interest. The legitimacy of adhering to agreements made is implicit in the fact that the agreement prospect, in order to be agreed to, will be consistent with individual self-interest. Granting this, consent is natural. There is no need to suppose that contractarians are committed to *denying* that one consents only to that which is in one's self-interest, given one's situation, the situation of the other bargainers, and the available options. This is absolutely consistent with Hume's assertion that our consent binds us *only* because of our *interest* in being thereby bound. Gauthier, too, recognizes that this is *not* incompatible with contractarianism.[75] Contracts are binding not because we are duty bound, as if independently of our interests, but because we have an interest in being thereby bound.

Hume believed natural sentiment is the grounding for why we behave morally. Hobbes, on the other hand, maintained that to be moral, i.e., to comply with one's covenants, is in one's own interests. And this is a very much different interpretation than what has come to be considered the standard interpretation of

---

[74] "They imagine not, that their consent gives their prince a title: But they willingly consent, because they think, that, from a long possession, he has acquired a title, independent of their choice or inclination" (Hume, "Original Contract," 457).

[75] David Gauthier, *Moral Dealing: Contract, Ethics, and Reason*, Ithaca and London: Cornell University Press, 1990, 56.

Hobbes. There is no "duty" then to stick to one's agreements unless we have an *interest* in doing so. The *raison d'etre* of contractarianism is precisely to show that we *do* have an interest in sticking to our agreements. After all, if we had an interest in making them, we have an interest in abiding by them. The duty comes once we accept a moral inclination and not before. The advantage of morality is, after all, indirect. To offset seeking *direct* advantages, we see the advantage of setting up a moral institution that *makes* it one's duty to keep agreements. So long as we recognize the duty is created, we must dismiss the interpretation that we have a duty to obey the natural laws found in the State of Nature. There is no *duty* to obey the laws of nature. Simply, the laws of nature point out what is *reasonable*, what will likely satisfy one's interests.

3.2 RATIONAL SELF-INTEREST

One of the important claims of contractarian ethics is that the guiding motivation of morality is rational self-interest. But by rational self-interest, contractarians do not mean that all people are naturally egoistical maniacs deliberately calculating their every action for their sole monetary profit with never a motive for others' interests.

One's interest may be the interests of another, as Butler rightly saw.[76] Thus, pointing out altruistic acts of people does nothing to harm the psychological claim that humans are motivated by their own interests solely. Likewise, it may be in one's interest to act in the spur of the moment; to be non-calculating. In golf, for example, the time for deliberate calculations is on the practice ranges and putting greens, not during one's backswing. Some calculations are appropriate; what club to use, what side of the pin to shoot for, whether to lay up. But during the set up and swing itself, that is the time to delegate all authority to the muscles; to keep the mind silent. Prospective love interests, as well, are best kept alive by acting "from the heart" rather than "from the head." Calculating how to be romantic is often to fail. None of these points detract from the claim that morality is (or can be shown to be) based on rational self-interest.

How then do contractarians interpret self-interest so that these points are not countervailing factors?

3.3 THE ATTRACTION OF SELF-INTEREST

[76] Joseph Butler, "Upon human Nature," from *Fifteen Sermons Preached at the Rolls Chapel*, London, 1726 [reprinted in Bernard Baumrin (ed.), *Hobbes's Leviathan: Interpretation and Criticism*, Belmont, Cal: Wadsworth Publishing Co., 1969, 16-25].

The attraction of basing morality on self-interest was evident as far back as the *Republic*. It is good to be moral, Socrates believed, because only in being moral does one's soul function properly: in harmony. There is nothing better for the self than to have a harmonious soul. That is a good to the self! The claim, after all, is not: "Get your soul in harmony because you're hurting others." The others seemingly do not matter. It is for the good of your own soul that you should be nice to others. The power of such an argument, if it succeeds, is to render morality a prudential course of action. No longer should we view morality and individual interest as conflicting. If the sole motivation to be "immoral" is that one can benefit from it, showing how this is actually a false premise, that immorality is a deficit and morality a benefit to the self, is to internalize morality as being the rational choice for all.

Socrates's argument is unlikely to be convincing however often parents try precisely this approach on their progeny. The earliest moral lessons parents place on their children is that "bad" acts -- defined merely by what their parents forbid -- will merit punishment of some sort. As it is typically in one's self-interest not to be punished, it is derivatively in one's self-interest to be "good." At a certain age, always younger than the parents anticipate, the children learn that some "bad" acts can either be committed without detection, or the benefits derived from them nevertheless outweigh the consequent punishment. Another possibility is that some

children regularly receive punishment for all sorts of acts for no other reason than the punisher is in one of those moods. In such cases, as the children are being punished anyway, they might as well take the benefit from acts otherwise deemed "wrong," for example, the exploitation of younger siblings for their own gratification. For this group, since punishment is inconsistent and forthcoming anyway, it is no deterrent for "immorality." In fact, it is possible that inconsistent and regular beatings increase the likelihood of immorality. In all these cases, parents quickly require a further justification for "good acts" than the avoidance of mere punishment. This stage may be delayed for a spell by simply increasing the punishment. Quite likely, some parents are unable to muster the sophistication of Socrates's argument ever, and have no other recourse but to up the ante until the children leave, quite likely prematurely.

But the sophistication of Socrates's argument is itself unconvincing. It requires the belief that we have souls in the requisite sense, and that these are organised in such a way that certain behaviours in the external world directly affect it. Such a belief is taxed when we consider that identical actions may be moral or immoral depending on the context. That I push John into the river may have been a good thing if it was to have John avoid a drunk driver. To distinguish that case from one where I simply push him in because I don't like him, the soul requires connections through the senses, understandings in terms of concepts embedded

in the physical world, as well as feedback loops. In other words, the tripartite soul requires sophistication to the degree where the soul loses any explanatory power at all. But Socrates's answer for why be moral is unsatisfactory not merely because it requires specious metaphysical beliefs to support it. Under Plato's analysis, a harmonious soul is when the reason rules over the passions and appetites. This is not a recipe for a harmonious soul in all circumstances. Nor can it all by itself resolve the issue of what or what not is moral. We do not say it is moral to steal diamonds so long as one does so with a calm, cool disposition. The calculating assassin is not more moral than a frenzied spouse who kills her abusive husband.

Exasperated parents may simply decree that a certain act is right just because it is. This is the antithesis of the self-interest approach. The claim is that morality is not in one's self-interest, and that that's just the point. Kant took this approach.

3.4 KANT

Morality as a motivator is often thought to be precisely *counter* to self-interest. Duty, accordingly, is a motivator as much as self-interest. A prime example is Immanuel Kant. Kant advocated a strict deontologic ethic without appeal to self-interest of any sort. Kantianism, thus, is a counter theory to contractarianism. If the contractarian self-interest claim can stand on its own, it is necessary to show that Kantian

deontologic theory is supported by self-interest after all. This is easily demonstrated.

Kant advocates an ethic based on duty. It is not enough that one obeys a duty; it is moral so long as one's motivation in obeying that duty is of the appropriate sort. The appropriate sort is that which does not further one's self-interest. It is as if the proper moral motivation for Kant is a begrudging one. This, however, isn't quite appropriate, since, presumably one begrudges the act because it conflicts with one's self-interests, and so far as self-interests enter the picture, the less moral is the motivation. The proper motivation, then, is not to have concern for one's self-interest at all. The contractarian claim is that this is not psychologically possible. That is just not the way we work. And I shall argue that Kant is not unaware of this psychological fact.

According to Kantian ethics, it is not enough to know that to be moral one must do one's duty. We also need to know what one's duty is. Is it to be charitable? Pursue one's talents? Refrain from reading pornography? As a proper theory should, Kant provides us with a formula for ascertaining the moral action in any given situation. The principle, or formula, is the Universal Principle. This, roughly, states that you should act only on the maxim by which you would have others act.[77]

---

[77] Specifically, "Act only according to that maxim whereby you can at the same time will that it should become a universal law" (Immanuel Kant, *Grounding for the Metaphysics of Morals*, James W. Ellington (trans.) Indianapolis: Hackett Publishing Co., 1981 [Originally published, 1785], 30 [Paul Menzer's (Berlin, 1911) *Konligliche*

If I am thinking of stealing money and am wondering if my duty forbids or permits this act, I need ask myself whether I would like it if everyone steals. And I think, among other things, that if everyone steals, people would not tend to leave things lying around. Things would be chained down, thus making it more difficult for me to steal.[78] I might also think anything that I possessed might be stolen, and this may seem a cost greater than the benefit I might derive by my stealing others' goods. In my calculations in deciding between a system where everyone steals and a system where no one steals, I would have to add to the "everyone steals" system the cost of locking everything down to prevent my property being stolen as well as the cost of interminable vigilance over my own belongings. Kant's claim, and I agree with him, is that stealing is unlikely to be a maxim I could consistently will on everyone. Hence, theft is found to be immoral by appealing to the universal principle: I have a duty not to steal.

Kant's principle is supposed to provide universal agreement about one's duties. It can do so only so long as everyone has property they value more than the property of others. But this is unlikely to be the case for everyone. People living in

---

Preussische Akademie der Wissenschaften, Standard pagination, 421]).

[78] I am not thinking of the case where "stealing" is in fact a right. Of course, where theft is "permissible, it is no longer "theft" in the proper sense at all. I ask readers, therefore, to substitute "theft" or its analogues with "taking the possessions of others without their consent" or its analogue.

squalor may not be concerned about protecting the crumbs, rags, and debris they happened to have scrounged out of the dumps. These people will place different weightings in their calculations than those Kant proposed. Willing that everyone steals may come out ahead of willing that no one steals.[79] Barring this, it is unclear whether recognition of one's categorical imperative is sufficient to motivate us to act on them. It is questionable whether my mere willing that others refrain from stealing is sufficient evidence that I will or even should refrain from stealing.

These observations are beside the point, however. The interesting thing in relation to contractarianism is that Kant's principle relies, after all, on self-interest.[80] We recognize what is our duty on the grounds of *individual interest* alone. The fact that *I* would not like it if my maxim were universalized is what is being appealed to: *my* interests according to the hypothetical scenario. Only when the calculation is favourable to my self-interest, will I agree to abide by duty, and not before. Duty,

---

[79] So long, anyway, that the calculations are taken at the current time-slice, or where the current time-slice is accredited more weight than future time-slices, since it is conceivable that under a "theft is permissible" system someone may go from rags to riches, and at that point may wish for a "theft is not permissible" system.

[80] Schopenhauer made this observation in 1840 in *On the Basis of Ethics*, E. F. J. Payne (trans.) New York: Liberal Arts Press, 1965, pt. II, §7, 89-92.

then, contrary to Kant's express claims, is shown to be grounded on self-interest; precisely the view of contractarianism.[81]

One may point out that Kant believed it is not self-interest, *per se*, that we appeal to when evaluating whether our maxims can be universalized. Kant thought it would be rationally inconsistent to believe otherwise in immoral cases.[82] According to Kant, the impossibility of universalizing an immoral maxim is not based on a contingent notion of one's preference structure, but upon an objective concept of rationality itself.[83]

---

[81] Gauthier observes much the same thing, and calls it, erroneously, I believe, a "subversive reinterpretation of Kant" (Gauthier, *Moral Dealing*, 110). I do not pretend anything of the sort here. Simply, I believe Kant fails to ground a deontologic ethic, and he went wrong because, by his own account, he appeals to self-interested motivations. It is a critique of Kant. It points out that Kant can not consistently ground his ethics without appeal to self-interest, as he expressly intends. More, it points out that his moral imperatives could only be grounded on contractarian lines. It is idle to claim that this is an "interpretation" of Kant. It is a criticism. He fails his own objectives. For our purposes, it is more than that Kant fails. This, by itself, does not support contractarian thought. The important point is that it fails precisely because his underlying justification points more toward a contractarian grounding of morality than his deontologic view.

[82] Kant stated that some acts will be rationally inconsistent; he did not assert that they would *all* be so (Kant, 32 [424]). This should not affect my argument here. For those acts that are not irrational, self-interest is the foundation for our not willing them to be universal laws. For those acts that, accordingly, are irrational to will to be universal, what I argue below applies.

[83] Kant, 34 [426]; 36 [428-429].

Two responses can be made. For one, it is not clear that Kant succeeds in his claim that reason has no practical role.[84] If Hume is correct, reason is purely instrumental to our passions.[85] If so, showing the rationality of not acting on a given maxim does not show that self-interested motives do not play a key role in that calculation. More critical, however, is that it is highly suspect that Kant succeeds in showing that there are purely logical inconsistencies in abiding by certain maxims. It might be claimed that "allowing stealing" is an inconsistency *by definition*. There is no such thing as consensual theft. But if we consider whether we would will everybody to take whatever they fancy, then there is no *logical* inconsistency here. Nor is there any logical inconsistency in claiming that any person should kill oneself if they want to and are suffering unendurably.[86] What do I care if others kill themselves. Would I want others to kill themselves if they are suffering as much as me? If I am considering it for me, the consistent thing would in fact allow it for others

---

[84] See Gauthier, *Moral Dealing*, 115.

[85] "Reason is, and ought only to be the slave of the passions; and can never pretend to any other office than to serve and obey them," Hume, *Treatise*, pt. III, sec. III.

[86] Kant's example does not add the clause "if suffering unendurably", and this by itself is suspect. As it is suicide we are dealing with in the example, we do not need to check to see if we think everyone should kill themselves whether they want to or not. And even if that is what we are supposed to be asking, what does it matter to the suicide victim, should he succeed?

too. Why should I be so heartless to prevent it? And neither is there any logical inconsistency if I want to harm others. It is true that if I will that everyone abide by this maxim, I must recognize the likelihood of being harmed myself. But there appears nothing that makes the desire to harm both others and oneself logically or necessarily inconsistent. Mass murderers who turn the gun upon themselves when the police are closing in, however pathological, are not necessarily illogical.

It would be disastrous for Kant if people have differing views on what the categorical imperative dictates in any particular situation. It can not be that one person believes people must be charitable according to the universal test, while another person conceives no such duty after applying the universal test. Different results from the same test would destroy the desired universality of morality that Kant (and contractarians) want. Thus, wherever there is a disagreement about whether a personal maxim can be universalized (as in the cases of suicide, charity, pursuing one's talents, and possibly even harm), it is incumbent on Kant to show that one side is wrong. If the wrongness can be due to illogicality, or internal inconsistency, then the charge of elitism or paternalism against Kant can be avoided. But so long as there is an underlying appeal to individual interest, no such inconsistency can be shown.

3.5 MORAL REASONS

We have seen ways in which parents fail to show that morality is a good thing for their children to oblige. Of course it is conceivable that it is a good thing for the parents that their children are moral, at least toward them. Likewise, it is conceivable that to live in a society that is moral is a good thing for their children. Our question is not, "Why should *we* be moral," but "Why should *I* be moral?" It cannot suffice to claim that it is in one's interest to be moral to avoid punishment. This does not speak to the possibility of avoiding punishment, or cases where the punishment does not outweigh the immoral act. Nor for that matter does it explain why some acts are deemed punishable while others are not. The reason for such a distinction can not resort to punishment without being stupidly recursive. And if there is a reason to punish some sorts of acts over others, then this should be the reason not to do those acts all by itself, without recourse to the punishment attached to it.

This latter remark is not precisely true without qualification. It is conceivable that it is in my self-interest to adopt a system where you and others would behave morally toward me. But in order to ensure that you comply, it may be rational to erect a system of punishment. Once this is set up, it is presumably neutral as to who the defector is. Thus, it is rational that I adopt a system that would punish me for defection. My reason for adopting the system is different than my reason for abiding by the system. I may abide by the system of morality merely out of fear of

punishment, although I can accept such a system so long as the punishment system protects me in other ways I deem profitable for me. In other words, the punishment commands cannot be arbitrary. They must meet my self-interest, *largely construed*.

It is roughly in this way contractarians claim that morality is based on self-interest. But what do I mean by "largely construed"? This is an important question that will, in fact, take up the bulk of my thesis. Self-interest is certainly not the *direct* basis of moral motivation. Or, alternatively, unconsidered self-interest is not the basis of moral motivation. If on a whim I wish to punch my brother in the face, acting on it cannot be moral on most accounts even though it furthers my self-interest in some respects. If contractarians mean by "self-interest" this notion of "direct" self-interest, or "unreflected" self-interest that includes mere whims, contractarians would agree that morality is not based on "self-interest" so defined. Contractarians believe self-interest is *constrained* by morality, but they argue that the rational motivation for accepting a moral system is itself grounded on longer-ranged, or "considered" self-interest.[87]

---

[87] See David Gauthier, *Morals by Agreement*, Oxford, Oxford University Press, 1986, 23, and Hobbes's *Leviathan*, xiv (75); as well as Thomas Hobbes, *De Cive*, Sterling P. Lamprecht (ed.) New York, 1949, 3.31 (57-58). Perhaps this is not far removed from Kant's understanding. What is in one's considered self-interest may well accord with what passes the universalizable test. What does not pass the test is not considered to be truly in the actor's self-interest.

Reliance on "considered" preferences in this way, however, makes room for a separate criticism. If whatever people seek they call good, then if they do not seek what is rational -- even what is good for them if they were rational -- then it can not be the case that what is good for them is the rational act. The contractarian claim is that since x better enables people to pursue their disparate goods, people rationally, if not in reality, will see x as a good. However true that may be, there will be a conflict for a person A who does not see the good of x in this light. If good is defined in terms of self-interest alone, then x will not be good for A. If it is not good for A, it is irrational for him to pursue it. Thus, contractarians, at this level, advocate a contradiction. X for A is both good and not good.

Simply put, reliance on "considered self-interest" as a grounding for morality is very much like saying: "Morality is based on self-interest, so long as your interests are moral." If this is what the introduction of "considered" self-interest amounts to, the contractarian attempt to ground morality on rational self-interest is pathetic. Before we address this problem,[88] we need to be more clear about what constitutes self-interest.

3.6 THE NATURE OF SELF-INTEREST

[88] In chapter 4 of this thesis.

The contractarian claim that men are by nature self-interested is often held to be refuted by the empirical fact that people frequently have sympathies for other people. It would appear that if morality is based on self-interest, then people agree to morality in purely calculating terms. But the calculating model of contractarianism does not do justice to the phenomenology of moral sentiment. If we help a stranger, it is not because we expect some recompense. We just are benevolent, and are so out of goodness itself, and not out of some long-range, utility preference. Versions of this complaint have been seen in Hume, Rousseau, and Locke, and more recently by Taylor, MacIntyre, and Sandel.[89] Butler noted a distinction, however, that circumvents this compliant. Butler argued that our capacities to feel and act for others' benefit is not necessarily counter to self-interest. Benevolence, for Butler, is the result of action motivated by "proper" self-regard.[90] "By acting merely from regard to reputation, without any consideration to the good of others, men often contribute to public good."[91] The word "often" here is troublesome, I admit. Is it the

---

[89] Michael Sandel *Liberalism and the Limits of Justice*, New York: Cambridge University Press, 1982. Alisdair MacIntyre, *Whose Justice? Which Rationality?* Notre Dame: Notre Dame University Press, 1987. Charles Taylor, *Sources of the Self: The Making of Modern Identity*, Cambridge: Harvard University Press, 1989.

[90] Butler, 20-21.

[91] Butler, 20-21.

case that sometimes self-interested actions do not lead to the public good, or is it to be taken more strongly: that self-interested actions sometimes are *counter* to the public good? If he allows self-interest to in fact inhibit the public good, it is difficult to maintain that morality can be based entirely on self-interest. And surely empirical evidence supports the claim that self-interested actions are often counter-productive to the public good. The existence of prisons attest to this fact. People are malevolent at times, and so Butler requires an explanation as to how this is possible if we are nevertheless designed by an omnipotent deity to be benevolent by merely abiding by our self-interest. Butler's explanation is that "true" self-interest will never run counter to the public good.

> There is a manifest negligence in men of their real happiness or interest in the present world, when the interest is inconsistent with a present gratification...thus they are often unjust to themselves as to others.[92]

So, when people are unkind or cruel to others as we see prevalent in the world, this is not due to any fault in God; merely in man. They fail to pursue or fail to recognize their *own* "true" good, and that is why they can fail to be benevolent.

---

[92] Butler, 25.

The main concern in this thesis is how to flesh out what one's "true" interests are. Butler's claim is that we need to attend to the voice of our conscience.[93] This will not help unless we know *a priori* what sort of a voice our conscience has. And there will be evident circularity if it turns out that we know what our conscience tells us not by introspection, but by post hoc assessment of the recommended actions. In other words, I fear that if the act is not benevolent, it is *then* judged not to have been in accord with our "true" conscience. What our "true" interests are may well be unrelated to what interests an individual might actually have. Discovering one's true interests by one's conscience becomes hopelessly circular. Butler avoided this circularity by what may be now termed a "communitarian" manoeuvre. His belief is not simply "one's interests are not truly one's interests unless they are benevolent." Rather, he believes (along with Rousseau, MacIntyre, Taylor, and Sandel) that we share interests. In particular, we share a common idea of the good. "Men are so much of one body, that in a peculiar manner they feel for each other," and "Mankind are by nature so closely united, there is such a correspondence between the inward sensations of one man and those of another."[94] Any interest in oneself that is *not* in such close correspondence to the interests of another will be regarded as perverse

---

[93] Butler, Sermons II and III.

[94] Butler, 22.

in some way. The pursuit of any wayward interest will have to be misguided. It is misguided not because they are doing evil to others, but because "they do evil to themselves, too."[95] And this is because we are, accordingly, "one body in Christ."[96]

I shall ignore Butler's reliance on religious speculations. Charles Taylor supplants the "one-body-in-Christ" thesis with an only slightly less religious model of an Aristotelian good that we all intuitively feel, if only we examined ourselves sufficiently. What is intriguing to note is that Butler is often cited as criticizing the psychological egoism prevalent in contractarian thought, whereas these reflections indicate that Butler's own theory may be viewed as advocating psychological egoism: Pursue your own true good, and that will necessarily benefit others. Given this, a normative theory would be: *Pursue your own true good always*. As this is compatible with contractarianism (so long as it is suitably interpreted), then the general interpretation that Butler critiques Hobbes for advocating psychological

---

[95] Butler, 23.

[96] Butler quotes from Rom. xii, 4,5.

egoism is ungrounded.[97] If people pursue *only* their own satisfaction, then benevolence will abound.

One may contest my claim that Butler is himself a psychological egoist on the grounds that "psychological egoism" is a thesis much more narrowly defined than simply saying people act in their own interests. Psychological egoism, for some, is the thesis that people never act with others in mind; never act to further moral motives; never act for reasons of benevolence.[98] Butler and Socrates do not say *merely* that being moral is good for the soul; morality is *also* good for others. And this is what is denied by the strict thesis of psychological egoism. For Butler, "people enjoy doing good for others."[99] If the strict thesis is the correct rendition of psychological egoism, however, then neither are contractarians in general nor Hobbes in particular psychological egoists. Hobbes's thesis is that men always act in order to satisfy their own desires. But "desires" for contractarians are simply

---

[97] "When Butler set himself to expose the fallacies of the `selfish' psychology of human action, he found admirable examples of them in some of Hobbes's analyses of the `passions', and he did the work of refutation so thoroughly that he has perhaps made the notion that there is nothing in Hobbes but this `selfish psychology' ... current from his day to our own," A.E. Taylor, 36.

[98] See, for example, Bernard Gert, "Hobbes and Psychological Egoism," in Baumrin (ed.), 109. [Originally published in *Journal of the History of Ideas*, 28, 4, 1967.]

[99] Butler, 22.

whatever desires the agent happens to hold, and this should incorporate any moral sentiment he wishes to act on; including benevolence. As Bernard Gert rightly decrees, "for Hobbes, it is simply a matter of definition that all voluntary acts are done in order to satisfy our desires. But...he does not deny that we can desire good for another."[100]

A possible complaint against the contractarian understanding of self-interested actions is that this weakened version of psychological egoism is vacuous. Has it been so weakened it makes no interesting claim whatsoever? To distinguish this weaker version from the strong version of psychological egoism, Gert calls the second "tautological egoism."[101] Any action, whether it produces acts good for the public or bad for the public or even good or bad acts for the agent herself, is said to be motivated by self-interest. But if the consequence of the act is not in the interest of the agent herself, for example driving while drunk and as a result crashing and being seriously injured, in what sense do we mean the act was motivated by "self-interest"? Butler appears to be right when he says some acts are "really" in our self-interest, while other acts are not. But the claim of tautological egoism all by itself can not seem to accommodate this crucial distinction. And

---

[100] Gert, 112. For example, Hobbes speaks of the natural affections of parents for their children (*De cive*, 3).

[101] Gert, 111.

without it, the contractarian hope of grounding morality on "self-interest," broadly construed, seems destined to fail.

The received answer relies on an understanding of what interests are *rational* to pursue, rather than what interests are approved of by one's conscience. What interests are rational to pursue, however, is itself open to much debate, and, as we shall see in chapter 4, is itself largely dependent on what counts as one's interests.

Chapter 4

CONSIDERED SELF-INTEREST

My claim in the previous chapter was that the only way to conceive of the ontogeny of morality is through satisfaction of self-interest. Human psychology is such that whatever is not in one's self interest will be unmotivating. Further, morality is an institution that preserves, rather than curtails, self-interest given the social situation in which we find ourselves. Contractarianism is a moral theory that gives a viable account of how morality is grounded on self-interest.

Interests, however, are not simple things. An individual may have competing interests. In such a case, although the pursuit of morality may further one of these interests, it curtails another interest that the individual also holds. To claim that morality furthers self-interest, then, is not precisely accurate. Morality furthers only *some* interests. A problem, then, for interest-based ethicists is how to decide which interests count as being conducive to morality, and which interests do not. The purpose of this chapter is to flesh out in more depth what contractarians mean by self-interest when they claim morality is grounded upon that.

4.1 FRANKFURT'S MODEL

Frankfurt distinguished two types of preferences: first and second-order preferences.[102] First-order preferences are preferences *simpliciter*. They are preferences for or to avoid objects or states of affairs. That I prefer to have ice cream rather than a peach is a first-order preference. But suppose I am concerned about my cholesterol level. Then it may be the case that I would prefer not to have the ice cream. This latter preference is not of the same type as the first. They are not merely competing interests, although the satisfaction of one is the disappointment of the other. Rather, the second preference has as its subject matter the first preference itself. It is not simply true that I prefer the peach. At the first-order level I do not. The second-order preference however gives commentary on the first. I would prefer not to have preferences for high cholesterol foods such as ice cream. I would prefer it if I would prefer peach.

The point of Frankfurt's model is that everyone will likely hold preferences that are vain and shortsighted at some time. Yet they may at the same time hold metapreferences, preferences about those shortsighted preferences. Given this, people may want social choices to be in favour of their metapreferences rather than

---

[102] H. Frankfurt,"Freedom of the Will and the Concept of a Person," *Journal of Philosophy*, 68, 1971, 5-20.

their first-order preferences. For example, I may have learned that acting on whims gives low odds of success. If a social choice rule helps prevent me from acting on whims, this may be seen as a good thing for me. The contractarian claim that morality ought to further individual preferences is no emendation for contractarianism if our individual preferences include debauchery, brutality, viciousness, and vice. Under Frankfurt's model, however, people could unanimously prefer that one's (first-order) preferences be overridden given that people have weak wills and recognize that they do. Incorporating into contractarian thought the emphasis on preserving second-order preferences over first-order preferences would thus solve the problem of whims. I can ask my crew to bind me and to refuse my orders while I sail past the sirens, and I trust that they will honour that reasoned second-order preference over my unreasoned first-order preference. I am better off that my crew decides to honour my second-order preference over my first-order preference. Analogously, we may be better off if social choice rules are based on furthering second-order preferences rather than first-order ones.

Let me be clear that I resist such results. It is questionable, for one, whether the distinction between orders of preferences is not mere fabrication. The claim is that second-order preferences have as their subject first-order preferences. But this is not obvious. My desire for ice cream is a first-order preference. It is not clearly the case, however well we can word it that way, that my second-order desire is

precisely *the not wanting to want ice cream*. More realistically, it is to be healthy. On this translation, we do not have different types of desires in the Frankfurtian sense; merely we have competing desires; one for ice-cream and one for health. There can be no practical guide to determine whether social rules should preserve first or second-order preferences.

Barring this worry, Frankfurt's model cannot help us unless we are content that second-order preferences always trump first-order ones. Contractarians cannot be content about this for at least three reasons. (i) For one, it is not the case that morality is based on or serves our second-order preferences. Morality in part defines them. That I prefer to bop Charles on the nose can only be circumnavigated by a higher order preference that is already normatively operative. Otherwise there is no grounding for why I should prefer one over the other. Why should I garner a higher order preference that prefers I do not commit preemptive violence? Assuming that people naturally have these higher order preferences seems empirically false. But even if they were naturally held by people at various degrees of cognizance, we still want to know why we should honour those rather than the first-order preferences they aim to supplant. We require a reason that appeals to the preferences we actually have. The contractarian claim that morality is founded on self-interest would be idle if self-interest "proper" is meant only those interests that are morally proper.

(ii) If the claim is that it is these particular second-order preferences that are satisfied by morality, an individual can still ask why should morality interest him then. The reliance on these second-order preferences become *standard-bound* rather than preference-based. (This is a topic I discuss at great length in chapters seven and eight.)

(iii) For that matter, it may be the case that I prefer not to worry about cholesterol. There seems nothing intrinsic in the distinction itself that second-order preferences should trump first-order preferences all the time. Think of an ordinary case. I want to ask Betty out on a date. Due to last minute jitters I catch myself preferring I did not have an interest in Betty in this way. Should this higher order preference win out here? Assuming that it should not, at least necessarily, we recognize that higher order preferences do not necessarily have greater weight than lower order preferences, and thus any rule to always follow higher order preferences when they conflict with lower order preferences is not a good guide.

There is either a rule for which preference order wins out, or the decision is left to the individuals themselves. If it is the former, then we have moved from self-interest, per se, and have adopted an external *standard* for determining what counts as "self-interest." But morality based on standards is not morality based on individual preferences, and hence contractarians cannot endorse it. If it is the latter, then there is nothing but the stronger preference that wins out, whether it be first

or second-order ones. And thus, as far as contractarianism is concerned, there is little use for the distinction. Morality cannot be bent to serve only one sort. And nor would it even be wise to do so, since some first-order preferences may be intuitively moral while second-order preferences may be intuitively immoral.

There is also the further problem of an infinite regress with Frankfurt's model. If second-order preferences trump first-order preferences, there may well be third-order preferences that trump second-order ones. Perhaps there are higher order preferences above these. Sincerely doubting the efficacy of a second-order preference cannot merely be the whiny complaint of the first-order preference the metapreference had intended to squash. If we can be at all serious in wondering about the efficacy of a second-order preference, this must be on the basis of a third-order preference, at least given Frankfurt's model. Alternatively, if we judge metapreferences by something *outside* our preferences (for example, according to duty, virtue, faith, or precedent) then we are simply abandoning our anchorage in individual preference. There is no way such a non-preference-based standard can legitimately be called contractarian.

4.2 GAUTHIER'S CONSIDERED SELF-INTEREST

Butler distinguished between true and false self-interest, and claimed that morality furthers only one's true self-interest. Which interests are one's true interests are

those in accordance to God's design, recognized by one's conscience.[103] David Gauthier, in his seminal work, *Morals by Agreement*, adopted the Butler-like lines of distinguishing between types of preferences. Rather than basing "true preferences" on God's design, however, morality for Gauthier was grounded on "considered" self-interest.[104] Gauthier argues that we can base morality on rationality. This is certainly more easily done (albeit illegitimate) if we introduce moral normative assumptions in our definition of rationality. I shall argue that this is what he does. A rational act, for Gauthier, is conceived of as that which furthers one's preferences -- *provided that one's preferences are "considered."* It is the introduction of "considered" preferences that I find problematic.

Rationality, for Gauthier, is identified with the maximization of utility, and utility is associated with preferences.[105] How we determine what one's preferences are, however, is a matter of some concern. Gauthier argues that preferences must be *considered* for otherwise "one may have reason to act contrary to one's actual

---

[103] Joseph Butler, "Upon Human Nature" in *Fifteen Sermons at the Rolls Chapel* [London, 1726] reprinted in Bernard Baumrin (ed.) *Hobbes's Leviathan: Interpretation and Criticism*, Belmont, CA: Wadsworth Publishing Co., 1969, 16-25.

[104] David Gauthier, *Morals By Agreement*, Oxford: Oxford University Press, 1986, 23, 29-38.

[105] Gauthier, 22.

occurrent preferences given that, were one adequately to reflect, one would change those preferences."[106] The problem is that this removes the conception of rationality from strict preference fulfilment. We need now ask *what* is the process of reflecting on one's preferences? What is the mechanism of weighing, of considering one's preferences? Under Gauthier's program, it can not be preference -- for that is what is being assessed -- and nor can it be rationality -- for what is rational is defined *once* the preference is considered. If we do insist on "considered" preferences, that is an admission that rationality is *more* than merely preference maximizing; it must *also* be involved in preference *ordering*. This precisely is the complaint raised by James Fishkin. "The criteria for considered preferences specify nothing about the appropriate conditions for preference formation."[107] And if this preference ordering is found to be *external* to the individual's subjective ranking, then we have an objective account of what preference it is rational to have -- based, for example, on the wisdom and experience of generations and *not* on the individual's own occurrent understanding. "Common sense," Baier claims, "allows the rational assessment of the relation between some beliefs and some attitudes....We do not

---

[106] David Gauthier, "Morality, Rational Choice, Semantic Representation: A Reply to My Critics," *Social Philosophy and Policy*, 5, 2, 1988, 192.

[107] James Fishkin, "Bargaining, Justice, and Justification: Towards Reconstruction", *Social Philosophy & Policy*, 5, 2, 1988, 54.

allow any and every outcome of reflection to be rational."[108] For example, if rather than the nail I hammer my thumb, most of us will assume this was not my preference. Typically, then, we do not assume preference from mere behaviour. Of course, one may conceive of some rational explanation for desiring to hammer my thumb rather than the nail. Perhaps I had wanted to avoid killing a butterfly that had alighted upon the nail. But if my "reason" was because a dog told me, we may think the *reason* irrational, according to Baier. We may be entitled to think it so given the community's belief that dogs do not talk and, even if they did, that is not sufficient reason to abide by what they say. Thus, on Baier's view, we have a concept of rationality that supersedes whether or not one's attitudinal preferences coincide with one's behaviourial actions.[109] We have, instead, "standards by which to judge whether the preferred accords with the preferable."[110]

The typical complaint against objective standards is that we have no rational basis for accepting them. An objective criterion for Baier, however, need not be

---

[108] Kurt Baier, "Rationality, Value, and Preference", *Social Philosophy and Policy*, 5, 2, 1988, 41.

[109] See Gauthier's discussion in *Morals by Agreement*, 27.

[110] Baier, 41.

fixed. It need not be floating in the "queer" metaphysical realm noted by Mackie.[111] Rather it is dependent upon the social norms of the times. Baier and Mackie are in fact closely related on this point. Mackie is often taken to be adamantly against any objective standards of ethics. This is simply not so. He openly admits that, "Evaluations of many sorts are commonly made in relation to agreed and assumed standards....Given any sufficiently determinate standards, it will be an objective issue, a matter of truth and falsehood, how well any particular specimen measures up to those standards."[112] Moreover, the selection of any particular standard is not arbitrary; it is based upon the function the standard is intended to achieve. If the function of morality is to reduce conflict, then any standard or principle that serves this is good. It is an objective fact, for the most part, whether someone satisfies this particular standard or principle. All Mackie denies, as well anyone should, is that the standard itself is objective. It is not an objective fact existing independently of men that morality must resolve conflicts, or that the purpose of morality is to make men happy, or to allow them to pursue their own goals. There is no threat to morality to deny the objectivity of these values. Denying their objectivity does not

[111] J.L. Mackie, *Ethics: Inventing Right and Wrong*, Hammondsworth, UK: Penguin books, 1977, 38-42.

[112] Mackie, 25-26.

make them arbitrary, however. It depends on the use we want for them. Basing morality on bicycle riding ability, for example, is not a use we have much call for.

This objectivity based on subjective but non-arbitrary standards according to function is also the notion of objectivity that holds for Baier. The criterion of rationality for Baier is the reliance on "the public system of reasons [that] have the backing of the experience and wisdom of many generations."[113] Until we deem that dogs can communicate to people, a dog's talking cannot be counted as a rational reason. Although it may not be contrary to my reasoning to hammer my thumb given my belief about the message from the dog, my *reason* itself may be judged to be irrational. "Irrationality, but not contrariety to reason may depend on the manner in which preferences are held."[114]

One problem for contractarians in accepting Baier's account is that if we judge actions simply according to "the public system of reasons," then why not also accept the morals that are backed by the community's beliefs. To say that morality is rational but that rationality is a social -- and not individual -- standard is to give up -- not answer -- the search for a rational self-interested account of why we should individually accept morality. If so, we have lost the individual appeal of

---

[113] Baier, 44.

[114] Baier, 42.

rationality and with it the individual appeal of morality. We cannot both introduce some *objective* standard for determining what counts as a considered preference according to external societal standards, and yet also claim that *utility* (preference maximization) is subjective.

One last comment against the introduction of considered preferences needs to be noted. According to Gauthier, "maximization imposes conditions on preferences."[115] Yet, "unconsidered preferences may meet, considered preferences may fail the conditions for maximization."[116] If considered preferences may fail conditions for maximization, then how can maximization measure considered preferences? And more telling, if considered preferences may fail the conditions for maximization, how can a maximizer rationally adopt the constraint of considered preferences? The introduction of "considered preferences" invariably puts one in a bind.

For contractarianism to succeed, a concept of rationality must be found that remains subjective; that can appeal to the individual's own motivational system. What is "considered" must be dependent entirely on the agent. David Schmidtz proposes such a model.

---

[115] Gauthier, *Morals by Agreement*, 38.

[116] Gauthier, 38.

4.3 SCHMIDTZ'S REFLECTIVE RATIONALITY

Like Gauthier, Schmidtz disparages the concept of rationality that makes it simply a means-end instrumental relation. As with Baier, Schmidtz holds that one's ends can be rationally or irrationally held. On the old instrumental model of rationality, what is rational is simply whatever achieves one's end. Thus, rational evaluation of one's ends is an impossibility. Schmidtz disagrees. Ends can be evaluated in terms of how they further further ends. As stated, this does not alter the means-ends instrumental model, since whenever we evaluate an end in terms of a higher end, we simply treat the first end as a mean. Schmidtz means more than this. Some ends are what he terms "maieutic ends." "Maieutic ends are the further ends for the sake of which we choose final ends."[117] Accordingly, there is a distinction between *pursuing* a final end and *choosing* that final end. One may choose a final end in terms of some further end, but this is not the reason why one pursues a final end. In the straight means-ends model, eliminating the end renders the mean as pointless. If I want to court Betty I should buy her flowers. But if I no longer want to court Betty, in fact have learned to dislike her, I should no longer want to buy her flowers. But with maieutic ends, "eliminating the further end is part of the process."[118] For example, once Betty decides to marry Barney rather than Malcolm,

---

[117] David Schmidtz, *Rational Choice and Moral Agency*, unpublished manuscript, ch. 3, "Choosing Ends," 4.

[118] Schmidtz, ch. 3, 7.

she has satisfied her desire to be married. But no longer having to make a decision of who to marry does not eliminate (although to be sure it lessens) her desire for Barney. Desires to be married, like desires for a career, are maieutic ends, Schmidtz reasons. Final ends can be rationally assessed in accordance to how well they satisfy maieutic ends, and particular maieutic ends can be judged by the light of one's overarching maieutic end. Although Schmidtz considers the possibility of other overarching maieutic ends, he posits the overarching maieutic end to be "having something to live for." This beats out mere survival because if there is nothing to live for one would not necessarily want to survive any longer. Suicide and euthanasia, then, have a chance of being considered rational acts.

There is one more element in Schmidtz's reflective rationality model. One of the problems with the instrumental model is that it leaves loose ends. We can justify acts according to how well they satisfy particular ends, but we cannot justify the ends themselves. They just sit there. If the beginning of the chain is not justified, it is difficult to see how the last link in the chain is justified. Lacking this justification, it is difficult for some to see how we can call the last link in the chain a rational act. Frankfurt allows that some ends are just settled.[119] The infinite chain complaint is mere philosophical sophistry, he believes. In arithmetic, for example, it is possible

---

[119] Frankfurt, 16-17.

that we have made an error in our calculation, and so it is wise to double check. But at some point, we stop double checking. Our final answer is not unjustified merely because we did not recheck one more time. Thus Frankfurt is not particularly bothered by the infinite regress argument. Some ends are decisively accepted in a non-arbitrary, wholehearted way, and this is enough. Not for Schmidtz, however. He justifies the maieutic ends by the particular pursuits themselves. In order to have something to live for, we need particular ends. But the particular ends are themselves justified by our needing something to live for.[120] Thus he offers a circular, but non-vicious, justificatory scheme of ends. "By closing the justification no loose ends need further justification."[121]

Schmidtz's plan was to show how even the narrowest of conceptions of rationality "has the resources to explain the rational choice of ends, and further, to do so without leaving loose ends."[122] The model breaks down, however, on both points. On the old model, the postulated maieutic ends are simply the final unjustified ends in question. The distinction that maieutic ends fall away like missile platforms on their satisfaction unlike final ends is misleading. If I do not want to

---

[120] Schmidtz, 3, 14.

[121] Schmidtz, 3, 15.

[122] Schmidtz, footnote 21, 3, 17.

court Betty, I will not buy her flowers. But likewise, if Betty does not want to get married, she will not marry Barney. Maieutic ends and final ends are thus not different on this score. And if I want to court Betty but only for one night, my buying her flowers may satisfy that. But I do not suppose that Schmidtz would want to call a desire for a one night stand with Betty a maieutic end merely on the grounds that the desire to buy her flowers fell from the corpus of desires like a missile platform.

Maieutic ends seem merely to be the wholehearted ends described by Frankfurt. They can be big or small, depending on the level of abstraction we give them. My point here is that recognizing the existence of maieutic ends does not alter the means-end model of rationality in any significant way. It could so long as these special ends are themselves justified in the non-vicious circular way that Schmidtz proposes. But his model breaks down here as well. Having something to live for, the overarching maieutic end, can be achieved by *any* means, and so does not justify the *particular* end Schmidtz needs to close the circle. That Betty wants to be married is satisfied by her marrying Barney, but it does not by itself *justify* her marrying Barney given that there were other options. Failing this, the justification of having something to live for is not satisfied by any particular act. Rather it is something one must grow into. Decisions are justified merely on the grounds that *some* decision needs to be made, or that *some* decision is better than *no* decision.

This fact is not grounds to justify the particular decision made. Failing this, the circle cannot close so neatly. And hence there will inevitably be loose ends.

4.4 TAYLOR'S STRONG EVALUATION

Charles Taylor also belittles the means-end instrumental model of rationality.[123] Or, if that is all there is to being rational, he suspects that morality must be linked to something higher. This something higher is autonomy.[124] Rather than merely following one's considered self-interest, a being must be autonomous. Being autonomous is more than being self-regulating. A squirrel, a schizophrenic, or a sociopath may be self-regulating. Whatever occurrent desires the agent has, as long as those are the ones he pursues for the duration of the time he has that desire is indicative of self-regulation. There is nothing in the concept of self-regulation that requires reflection on what it is one is regulating. Frankfurt believed that critical reflection is the essential ingredient for personal autonomy.[125]

---

[123] Charles Taylor, *Human Agency and Language: Philosophical Papers*, Vol I, Cambridge: Cambridge University Press, 1985, 17.

[124] Taylor is not the only philosopher that conceives morality and autonomy as being necessarily linked. See also John Dewey (*Theory of the Moral Life*, New York: Holt, Rinehart and Winston [1908] 1960); Lawrence Haworth (*Autonomy: An Essay in Philosophical Psychology and Ethics*, New Haven: Yale University Press, 1986); Diana Meyers (*Self, Society, and Personal Choice*, New York: Columbia University Press, 1989); and Gerald Dworkin (*The Theory and Practice of Autonomy*, Cambridge: Cambridge University Press, 1989), to name a few.

[125] Frankfurt, 7.

To be self-regulating, all one needs is to choose actions that will satisfy one's desires. But to capture what we mean by autonomy, also necessary is the ability to reflect on one's desires, to decide what goals one ought to pursue, to decide whether some desires are worthy of actualizing or not. Although Frankfurt's notion of self-reflection is necessary for autonomy, Taylor believes it is not sufficient. Reflection on whether to abide by one's desires may be satisfied in two ways, a weak way or a strong way. Weak deliberation is such as to see which of two desires one should satisfy now. For example, one may desire to swim and eat, yet be unable to do both simultaneously. Thus, one needs to decide which of these two occurrent desires one should satisfy first. This sort of deliberation about one's desires is something a squirrel should be capable of doing. If our intuition that squirrels are not autonomous is correct, then some further distinction is required. The further distinction, according to Taylor, is *strong evaluation*.

One engages in strong evaluation if one deliberates whether one should have that particular desire at all. Of course, the desire to swim or to eat need not be denigrated by a strong evaluator. What might not be tolerated by a strong evaluator, however, is the desire to smoke. A weak evaluator may have to decide whether he should smoke now or swim now. It is conceivable that after the swim, he will smoke. A strong evaluator, on the other hand, would consider whether he

should smoke *at all*. Other sorts of desires that require strong evaluation include the desire for extra-marital affairs or the desire to injure those who irk us. These are sorts of desires that an autonomous being would deliberate about in the strong sense, and not just in the weak sense. An autonomous being would decide not merely whether he should have an extra-marital affair now or later, but whether he should have an extra-marital affair at all. A strong evaluator is engaged in a "deeper" level of self reflection than in deciding between whether he wants to chase an occurrent desire now more or less than he wants to chase another occurrent desire now. It is not sufficient for the autonomous being (a strong evaluator) that he thereby pursue a desire merely because he has it and it is not contingently in conflict with any of his other desires. And this, accordingly, is not something the weak evaluator would consider. The strong evaluator may reject outright the pursuit of a particular desire. The ability to make such decisions is the mark, for Taylor, of the autonomous being.

It is debatable at this point whether an autonomous being need to decide the issue one way or the other. It is conceivable that after strong evaluation he may decide to have extra-marital affairs, smoke, and hit those who irk him. Once that has been decided, then further weak evaluations about this matter may crop up. He is autonomous not to the extent that he always is engaged in strong evaluation, but

simply that he has engaged or is capable of engaging in strong evaluation about his first-order desires.[126]

For those who argue for the autonomy-morality connection, they must explain why autonomous beings will necessarily decide in favour of moral choices. Taylor's response to this is that strong evaluations are made on the basis of one's self-identity. One's self-identity is in turn constituted by one's community.[127] But then it would seem a simple (too simple) step to connect autonomy and morality. If one's community happens to disapprove of extra-marital affairs, then presumably you yourself do as well on this view. Thus, after strong evaluation, one will conclude against extra-marital affairs. Whatever is considered moral in your community will be what an autonomous being will choose. This relies on an excessively strong connection between self-identity and community values. As I have already discussed in chapter 2, I cannot deny that we are influenced by our community. If our strong evaluations (both the outcome and the perceived need for making a strong evaluation in a given situation) are dependent on what sort of personal

---

[126] It is not clear whether Taylor agrees with this. This gives the picture that an autonomous individual is not constantly engaged in strong evaluation. Taylor, however, claims that "the kind of being we are to realize is constantly in question" (Charles Taylor, "Responsibility for Self," in A. Rorty (ed.), *The Identities of Persons*, Berkeley, University of California Press, 1976, 289). This may overstate the case.

[127] Charles Taylor, *Sources of the Self: The Making of the Modern Identity*, Cambridge, Mass: Harvard University Press, 1989, 35-39.

identity we have, then it follows that our strong evaluations (which occur in both interpersonal and intrapersonal matters) will be influenced by these external social factors. We need not deny that social influences greatly effect our strong evaluations to deny that we should expect any homogeneity from this fact. The reason is, our identity is constituted by a multiplicity of social variables. That we are socially influenced, or even socially constituted, is not to say we are influenced or constituted by only one thing. The term "social" is a catch-all term for a multiplicity of factors, including our social positions, our career choices, upbringing, surroundings, peers, associations, keep-sakes, etc. Although there is some plausibility that any one factor will have an influence on many people, it goes against probability that many individuals will be equally affected by all the same variables together. So although we are socially affected (or stronger: constituted) it does not follow that we will share the same values. We should not expect homogeneity of moral beliefs among individuals simply on the basis of a common set of social factors. We are influenced differently by different factors within that set.

Barring the questions we need to raise about the implications of the strong evaluation thesis, we may decide to raise questions about the strong evaluation distinction itself. For example, there is one sense in which this is no augmentation of Frankfurt's model. Frankfurt defines a wanton as one who lives almost entirely

by following first-order desires.[128] Nevertheless, this wanton is able to decide which desire to achieve at any given moment,[129] for surely it is impossible to have only first-order desires that never even contingently conflict (unless one has only one desire). According to Taylor, however, evaluation between contingently conflicting desires is the mark of a weak evaluator. On Taylor's analysis, nothing but the very minimal life forms could live entirely on first-order desires. It is evident, then, that what Frankfurt calls a wanton, Taylor would call a weak evaluator. But once we recognize this difference in vocabulary, we see that there is no difference in their recognizing the same conceptual distinction between having a desire, deciding between desires, and deciding about one's desires. If there is no relevant distinction between Taylor's and Frankfurt's models, then the same criticisms that have been raised against Frankfurt apply equally well to Taylor.

A further problem with Taylor's model, however, is this: Part of Taylor's motivation for introducing strong evaluation is that Taylor's distinction between first and second-order desires cannot adequately address how second-order desires originate. It seems possible that second-order desires may themselves have been un-autonomously accepted. Frankfurt believes there is a point where the

---

[128] Frankfurt, 11.

[129] Frankfurt, 11.

satisfaction of our second-order desires "resonates" through us, and we need no higher order desires evaluating these any more than we need to continuously recheck our arithmetic.[130] This "loose end" bothered Taylor, evidently, as much as it did Schmidtz. Moving from a model that posits desires judging other desires to a model that has people "evaluating" desires is supposed to eliminate this reliance on settled preferences for autonomy.[131] It does not, however. This same problem (if it is a problem) remains. That I vow never to entertain a particular desire is itself an evaluation that will just seem settled to me. That is, it will be based on the type of person I am (or want to become), and this personal identity must merely be that with which I do in fact identify.[132] If we ask, but *ought* we identify with this, we meet precisely the infinite regress problem raised against Frankfurt's model.

There is a fundamental problem with linking morality to autonomy (rather than rational self-interest) that tarnishes all attempts. We can concede a strong link between morality and autonomy and yet deny that this is enough to support the

---

[130] H. Frankfurt, "Identification and Wholeheartedness," in H. Frankfurt, *The Importance of What We Care About*, Cambridge: Cambridge University Press, 1988, 166.

[131] See Gary Watson, "Free Agency," *Journal of Philosophy*, 72, 8, 218.

[132] Taylor wants to say no strong evaluation is settled, since any strong evaluation is open to re-evaluation (Taylor, *Human Agency*, 38; *Sources*, 20). But this is beside the point here. Just as one may recalculate 17 x 82 at some future time, the limited double checking that goes on now reaches a non-arbitrary degree of decisiveness.

claim that being moral is dependent on being autonomous. Defendants of the autonomy view also need to show that the strong evaluations of an autonomous individual will necessarily be moral resolutions. It is not at all clear that whoever is autonomous (a strong evaluator) is necessarily moral. Whether an autonomous agent is engaged in deliberating about moral matters, deliberating about moral matters is not identical to being moral. A thief may have deliberated about his desire to steal; simply he resolved that it was a "good" thing. What matters if we find a link between autonomy and morality if that does not give us any guidance as to how the autonomous person ought to resolve particular moral issues?

What else is required is (i) the claim that to be autonomous is, in some sense, to lead a well-adjusted life, and (ii) the claim that a well-adjusted life for social beings will likely include some notion of interpersonal skills. One needs to be socially well-adjusted in order to be intrapersonally well-adjusted. If so, since part of our notion of being socially well-adjusted incorporates the moral domain (interpersonal relations), it would seem to follow that part of being well-adjusted will be to be moral. Unfortunately, this would not yet equate to saying that an autonomous individual will necessarily prefer a particular moral action. Although a moral character is the disposition to do the right thing at the right time, this does not mean necessarily doing the right thing in a particular case. Other factors need to be accounted for: including cognitive factors (knowing that a case for strong

evaluation is at hand is dependent on a variety of cognitive skills) and psychological factors (being alert enough, not run down, sick, tired, disinterested, preoccupied, in a rush, etc.).

Even if it is shown that an autonomous agent always acts morally, it requires further argumentation to show that non-autonomous agents do not generally act morally. If autonomous and non-autonomous agents are equally capable of acting morally, then the concept of autonomy is irrelevant to morality. Claiming that acting morally is one of the conditions of autonomy is to simply beg the question. Nevertheless, this, in effect, is precisely what is argued. A criterion for a minimal level of autonomy is based on our notion of what a normal individual is. Part of our notion of a normal person will be that he is a moral individual to some extent.[133] It is normal, we think, that an individual has certain moral sentiments. Morality then slips into our defining concept of autonomy. Under such a view, the link between autonomy and morality is not based on any theoretical account, but simply on descriptive folk-psychology. This is not necessarily bad. That autonomy becomes

---

[133] For example, Haworth and Groarke claim the following: "Since we *are*, whatever else, moral selves, a life which denies this by failing of reflective morality to that extent is nonautonomous. It is nonautonomous because it is insensitive to the sort of person one is." (Lawrence Haworth and Louis Groarke, "The Relation Between Autonomy and Morality." Paper presented to the Canadian Philosophical Association Meeting, Montreal, 1995, 12.) Not only does this approach beg the question of whether morality and autonomy are linked, it merely shows a contingent connection between morality and autonomy. In the Hobbesian state of nature, since our social sphere is one of amorality and war, autonomy will endorse preemptive violence. This seems peculiar.

relative to conventions (conventions of normalacy will presumably differ in different societies and times) should not surprise us. The problem, however, is in fleshing out these normal moral sentiments. Because people differ substantially and because many factors beyond disposition influence our behaviour, the "normal" moral sentiments will not likely be conceived in terms of any particular moral substantive issue. The test cases can not be such things as "helping this injured woman," for many "normal" people may fail to do this.[134] More likely, the "normal" moral sentiments will be reduced to something abstract. For example, a "normal" moral sentiment may be the having of "other-regarding" interests. As soon as we identify what a normal sentiment consists in these abstract terms, however, it is unlikely it will of its own prove to be sufficient for being moral. The question we have then is on what grounds does being autonomous necessarily enable one to

---

[134] See B. Latané and J. Darley, *The Unresponsive Bystander: Why Doesn't He Help?* Englewood Cliffs: Prentice-Hall, 1970. J.M. Darley and B. Latané, "Bystander Intervention in Emergencies: Diffusion of Responsibility, *Journal of Personality and Social Psychology*, 8, 1968, 377-383. Also interesting are J.M. Darley and C.D. Batson, "`From Jerusalem to Jericho': A Study of Situational and dispositional Variables in Helping Behaviour," *Journal of Personality and Social Psychology*, 27, 1973, 100-108.
and M.J. Lerner and D.J. Miller, "Just World Research and the Attribution Process: Looking Back and Ahead," *Psychological Bulletin*, 85, 5, 1978, 1030-1051). Accordingly, the more you believe that the bad suffer and the good are rewarded on earth, the less likely you will offer help to those suffering, since they must obviously "deserve" it. Recall Dante's pushing back from his boat those grieving in Hell since divine justice ordains they suffer without respite (Dante, *Divine Comedy*, *The Inferno*, Canto VIII). Here although strong evaluators are making moral decisions, the outcome is not clearly the normal "moral" response.

comply with the mores of one's society? Particularly, we return to the question, what purpose does it serve the autonomous individual to be moral? And this question asks what is the benefit to the individual in being moral. We are back at the need to define morality in terms of rational self-interest. The excursion to autonomy is therefore idle.

My point here is that the key of morality cannot be dependent on autonomy. Nevertheless, there is an obstacle in effectively resolving this debate. Autonomists and self-interest theorists have a different understanding of the function of morality. If morality's purpose is to have individuals lead good, virtuous, full lives, then autonomy and morality will clearly be interconnected concepts. In fact, the goal of morality or social institutions will be precisely to raise autonomy levels of individuals, since full lives is dependent on being autonomous. Morality on this view is broadly conceived; it encompasses every aspect of an individual's life, both interpersonally and intrapersonally. If, instead, we view the function of morality as controlling merely the interpersonal domain, then morality is seen merely as an enabler, a device in place to secure against interference in the pursuit of people's lives as the individuals themselves deem fit. How that life is led is left up to the individual. There is no antecedent notion of what a good life is by which his life is measured, at least within the intrapersonal sphere. Autonomy plays a part here in how full that individual's life can be, but the part it plays on this narrow view is an

amoral one. On the narrow view of morality,[135] autonomy may be a factor in achieving one's preferences, or certainly those preferences one autonomously endorses, but this is an amoral concern. Morality takes place within the interpersonal sphere and can be defended, so preference-based contractarians argue, by appeals to unmitigated occurrent preferences alone.

4.5 SUMMARY

I have argued that the criterion of what is or is not a considered preference is either an objective standard -- external to the individual actor -- or that it is decided wholly by the actor herself. If it is objective or standard-bound, then we have lost the motivational force for an actor to be "rational" when her preferences dictate some alternative choice. If morality is to be based on rationality, and rationality is pre-determined by some objective criterion of what preferences are *worth* pursuing, we have introduced a normative concept into the definition of rationality. This would be fatal to the contractarian goal of linking morality to individual self-interest.

On the other hand, if the determination of considered preferences is left wholly up to the individual, then we must admit that a considered preference can be none other than a preference held presently by the actor -- whether or not it will

---

[135] Defended in chapter 2.

sustain sufficient reflection. If this is true, then the introduction of "considered" preferences is wholly idle and we have, in fact, no different an account of rationality than that which furthers one's preferences.

Thus, introducing considered preferences as a criterion of rationality is either fatal or idle. We can do well to dispense with it. Having dispensed with considered preferences, we are left with a bare account of rationality that is contentless. What is rational is simply that which furthers one's preferences. Any preference. This is how the economists conceived it. What precisely counts as a preference when there appear to be many and when these conflict (as they often do) does pose a problem to the instrumental concept of rationality. The seeming resolve by considered preferences does not work. How a model using occurrent preferences fares will be seen in the subsequent chapters. We must first rid possible misinterpretations of the occurrent preference model. This is undertaken in chapter 5.

Chapter 5

OCCURRENT PREFERENCES

5.1 OCCURRENT PREFERENCES

Schmidtz, Baier, and Taylor found the inevitability of loose ends debilitating for the instrumental model of rationality. It need not be. Schmidtz is right in so far as pointing out that many of our ends can themselves be justified according to some larger end that we have. This should suffice to make us able to judge the rationality of particular ends. If we go back far enough, we will not find justification, per se, of our abstract ends. For the most part, I believe we can nevertheless be content with these far reaching or more basic ends. That I would like something to live for should be innocuous enough not to need justification. The problem with it, if there is a problem, will unlikely be found at this abstract level. The problem, if there is a problem, will be found in how I attempt to realize that goal. Wanting more to live for is not necessarily satisfied by becoming a thief, especially if one spends much time in prison as a result. The problem is not in the abstract goal of wanting something

to live for, it is in the execution of that goal. I agree with Frankfurt, then, that "loose ends" are not a problem, either for models of rationality or models of morality.

We have yet to solve the problem facing considered preferences, however. If preferences are subjective and relative, on whose criteria can we distinguish considered from unconsidered preferences? As Baier noted, "There does not seem to be any genuine empirical way to determine when someone has sufficient experience or has reflected sufficiently to see whether or not he has a fully, or even sufficiently, considered concrete preference."[136] No one other than the agent herself can determine whether her own preference is sufficiently considered. Gauthier, importantly, recognized this. "For if it be agreed that values are subjective, then there is no ground for appeal beyond what a person acknowledges." And if this is true, then we can not accept the qualification that Gauthier adds, "given that she reflected sufficiently and is fully experienced."[137] It is this qualification that introduces an objective or standard-bound imposition on what it is to be rational -- independent of the person's own values. Should her values be not sufficiently

---

[136] Baier, "Rationality, Value, and Preference", *Social Philosophy & Policy*, 5, 2, 1988, 39. We might add to Baier's quote, "at least prior to the action." Afterward, we may deem our previous preference was not sufficiently reflected upon. But I am solely concerned with determining a preference prior to an action. I discuss *post hoc* deliberations below.

[137] David Gauthier, *Morals by Agreement*, Oxford: Clarendon Press, 1986, 34 (both quotes).

reflected upon or based on insufficient experience, they are her values, nonetheless.[138] If values are subjective, then it must be left to the actor to decide on the degree of consideration she has given to her preferences. If we leave it to the actor, only she will be the qualified judge on whether she has given the matter sufficient reflection. As outsiders, we cannot circumscribe her actions based on *our* ideals of considered preferences. An implication of this strict preference-based model, the *considered preference* model, is that if she does not make any further reflection, then it is precisely because she assumes (or does not think otherwise) that she *has* sufficiently reflected on the matter. Given this, so long as she acts on a preference, it is symptomatic that she deems that preference to be a sufficiently considered one. In thinking that it is considered enough, the phenomenological account of that preference is that it *is* considered. By this account, every preference of hers would be considered (to her) -- no matter how "unconsidered" they may seem to someone else or even to the same person at a later time. The only way, then, to conceive of "considered preferences" as a meaningful distinction is to base

[138] According to Gauthier, "value is the measure of considered preference" (*Morals by Agreement*, 33). It may be true that I prefer to smoke yet value my health. Hence, the prima facie preference to smoke is not one's considered preference that "involves the endeavour to maximize value" (33). But there are instances where value is clearly *not* the measure of one's considered preferences. I value great music. I may even prefer to play the piano. But all things considered, I do not prefer to take the time to learn. Thus, although I value piano playing, my considered preference is not to play.

the criteria of consideredness on external factors. This is a problem for the considered preference view. One must either abandon preference-based ethics and adopt standard-bound ones, or abandon the reliance on considered-preferences in favour of a model permitting unrestricted preferences. I opt for the latter.

Although an individual may change her mind about how considered a previous preference really was, this does not unite a subjective account of preference with considered preferences. Gauthier insists "that the preferences be held in a considered way, and this is not a matter of what considerations the person cares to give, but what would in fact survive adequate experience and sufficient reflection" -- which he admits to be "vague standards but not dependent on individual whim."[139] They are certainly vague standards but they do not succeed in severing individual whims from what may count as sufficient reflection. It is true that I may prefer not to act on whims, and this preference may itself be derived from experience -- experience that showed me that whimsicality rarely satisfied my preferences. Alternatively, I may prefer to act on whims more than acting on the choices that my staid reason dictates. According to zen masters, for example, spontaneity reveals one's true preferences more than one's culturally influenced

[139] David Gauthier, "Morality, Rational Choice, Semantic Representation", *Social Philosophy and Policy*, 5, 1988, 194.

deliberative practices.[140] Here, especially, one's preferences are revealed in action rather than attitude -- which Gauthier admits is perfectly possible.[141]

Of course, if considered preferences -- those that survive experience and reflection -- could be revealed through whimsicality, this does not say they are *dependent* on whimsicality, I admit. But if considered preferences are denied any tie to individual whim, then the concept of considered preferences can *not* be subjective. If subjectivity means that the individual is the judge of what counts as true or real or considered for her, then if the individual decides to act on a whim, whimsicality and subjectivity are linked and can not be severed by the agent -- at least at the time of acting. If it is one's preference to act on that which one might later regret (and term it as mere whim), it is nonetheless rational to act on it -- for that is one's preference. This is in direct opposition to Gauthier's conception of rationality. He believes one is rational only so long as one furthers one's considered preferences. "If we may suppose that her revealed preference would not survive reflection unopposed, then we may agree that her choice is not fully rational."[142]

---

[140] I equate "spontaneity" with "whimsicality" although the connotation of whimsicality may make it analogous to that which is more flippant, more idle, more shallow than the connotation that "spontaneity" gives.

[141] Gauthier, "Morality, Rational Choice, Semantic Representation", 191.

[142] Gauthier, 34.

Given that one has a preference, whether it is based on mere whim or studious reflection, and one wants to further that preference, and it is rational to do that which one wants, then it is rational to further that preference -- considered or not. This must be the case so long as rationality is strictly instrumentally conceived.

From this we see why Schmidtz and Baier and Taylor and Gauthier wish for a more reflective model of rationality; a model that can critically assess ends as well as means. I have argued above that these attempts fail. They either offer no emendation to the standard instrumental model or they incorporate normative assumptions that do disservice to the goal at hand; grounding morality on rational self-interest.

An objection to my dismissal of these notable attempts can arise from observing how people decide between two or more preferences that are mutually exclusive. In such cases, one must determine which preference, or end, one wants to satisfy the most. One would expect that if we are able to make such choices, then we must be able to critically reflect in ways more noble than the bare means-end model allows.

This is not so. That one has to choose between preferences is simply tantamount to saying it is rational to satisfy that which one prefers and most rational to satisfy that which one prefers most. And this need not entail the introduction of *considered* preferences in the requisite sense. It is something even Taylor's weak

evaluator has the capabilities of achieving. In debating between the desire to smoke, say, and the desire not to smoke, one's action (smoking or not smoking) will be based on that preference (whether first or second) which takes precedence -- that preference which "wins out." This "winning" preference (again, regardless of whether it is a first or second-order preference) is the occurrent preference. To decide, for example, that one will not smoke will not necessitate that the desire to smoke will never reappear in the agent's list of preferences. And that is why the emphasis on occurrent preferences is all that we can admit. I call this the *occurrent preference model* of rational self-interest.

5.2 HEDONISM AND DEATH

My account of self-interest is hedonistic to the extent that the judge of desire fulfilment is the agent herself. Since the notion of hedonism carries with it some sense of advocating debauchery, let me clarify the hedonistic aspect of my theory as being phenomenologically dependent. This means the same thing. What counts for me as preference satisfaction is my belief or feelings at the phenomenological level that my preferences or aims have been satisfied. We make this distinction since it is possible that my aims have not in fact been satisfied; I might merely believe, erroneously, that they have been. A phenomenologically independent model of preference satisfaction may be called, after Parfit, a *success* theory. Aims

are satisfied if and only if they have been satisfied in fact, whether or not the agent is aware of it. It is rational on the hedonistic model for a philosopher to pursue a thesis that he *believes* to be sound, even if in actuality it is not. A non-hedonistic account will claim that philosopher irrational, since he has not in fact achieved his goal, which we assume is to defend a sound thesis, rather than to merely believe that his thesis is sound. The standards of non-hedonistic versions necessarily lie with a judge external to the agent, and are for that reason dismissed on my account. Parfit disapproves of the hedonist version of self-interest. Hedonist preferences concern only those features of our lives that are introspectively discernible. Since we assume the death-state is not introspectively discernible, no one can have preferences about events or states of affairs after one's death. Since this is patently false (we write wills and name beneficiaries in our life insurance policies), the hedonist account of self-interest must be false.[143]

To maintain a hedonistic version of self-interest, then, it is incumbent on me to explain how it is nevertheless rational to have eschatological concerns given nihilistic beliefs about the death state. It is quite easy, really. One's concerns on the occurrent preference theory are not concerns of immediate occurrent *satisfaction*. Simply, they concern present *aims*. I may have an aim now about what will happen

---

[143] Derek Parfit, *Reasons And Persons*, Oxford, Clarendon Press, 1987, 495.

tomorrow, and I will therefore, other things being equal, be rational to pursue that aim, even if there is a chance I will be dead tomorrow. That is, rational acts are not defined by satisfaction, but by anticipated satisfaction. Thus I need not ever actually satisfy, in the phenomenological sense, my aims to rationally pursue them.

This is all very well for aims that *might* be curtailed by death, someone might object, but if the aim is *conditional* on death, this is a different matter altogether. The satisfaction, phenomenologically speaking, is impossible. Thus even aiming at it is irrational.

I must remind my interlocutor that our aims are not restricted to self-satisfaction only. Parfit admits as much. "I could be purely self-interested without being purely selfish."[144] For example, my love for family and friends affect what is in my interests. So being self-interested does not deny I will act in the interests of others, so long as doing so is not against my interest. This shows that acting for interests other than your own is not irrational, so long as it is not contrary to your interests. Since nothing that occurs after your death can be contrary to your death-state interests, then arranging for any satisfaction of interests after your death will not be irrational on the occurrent preference model *at least from the perspective of your death-state interests* (which we assume to be nil). What then decides what

[144] Parfit, 5.

post-mortem planning is rational and what not? One's *present* aims. If you see that you do not need your wealth, and you care enough for someone else or some organization, then you are free and rational to arrange that this beneficiary gets your possessions when you know you will no longer desire them. We do not, then, have to give up the hedonistic understanding of self-interest-satisfaction to explain actions the satisfaction of which can occur only after one's death.

5.3 TIME INDEPENDENCE

Given that preference satisfaction is phenomenologically dependent and rationality is viewed strictly on an instrumental model, preferences will necessarily be *time independent*. This is important.

Parfit defended a version of the self-interest theory that is in certain respects like mine. He called it the *Present Aim Theory*. According to it, we should each do what will best achieve our present aims.[145] "In the case of reasons for acting that are based on value-judgements, or ideals, a rational agent must give priority to the values or ideals that he now accepts."[146] The emphasis on our *present* aims, coincides with my emphasis on *occurrent* aims or preferences. The occurrent

---

[145] Parfit, 92.

[146] Parfit, 155.

preference theory of self-interest is restricted to preferences held in the present. Thus, neither my model nor Parfit's will demand that we value a long term goal over a present goal. The motivation is the same for Parfit and me: we want a completely reductionist system; thus we cannot demand more of our subjects than they are willing to act on. If a present aim has more weight to the agent at the time of acting than some aim that cannot be satisfied until the future, there is nothing that requires the agent to forgo the present desire in favour of the future one. Both present aim theorists and occurrent preference theorists ask: "Why should I give weight *now* to aims which are not mine *now*?"[147]

The occurrent preference theory of self-interest is restricted to preferences held in the present. Preferences held in the present are not the same as preferences that demand satisfaction in the present, however. We have seen this when we considered the rationality of forming and pursuing preferences the satisfaction of which could not be achieved until after one's death. There are other sorts of preferences we form in which we expect only future gratification. At times, these "present-future preferences," as I shall call them, may conflict with desires the satisfaction of which are more immanent. Parfit is right to emphasize that there is nothing inherently rational that we should always forgo our near-sighted interests

[147] Parfit, 95.

at the expense of future interests, but he understands "future interests" not merely as interests one has now that will not be satisfied until the future. Rather, he views these as interests one does not have now, but which one might have in the future. Let us call these "future-present interests." He, like me, rejects our putting any weight on such kind of things. But we often have preferences now the satisfaction of which will only be in the future. These present-future interests, on mine and Parfit's theory, are part of the compendium of one's occurrent preferences.

It is a misconception, therefore, to believe these theories claim that it is necessarily rational to act on present urges that make longer-range plans difficult or impossible. If the longer range plan is one of the present aims or occurrent desires, then it is not clearly rational to forgo that. One problem for this view, then, is an apparent lack of prescriptive assistance. Nothing in the occurrent preference model is going to decide one way or the other over competing desires. The decision is a subjective matter concerning the agent alone, or so I claim. This is wholly unsatisfactory to many people, especially those whose intent it is to ground normative ethics on rational self-interest. If my model of rational self-interest cannot decide between competing occurrent desires, how is it going to advocate morality over immorality? This evidently was a worry of Parfit, and succumbing to this worry, he took a route very different than my plan here.

Parfit distinguishes three varieties of the present-aim theory: an instrumental version; a deliberative version; and a critical version. The instrumental version coincides with my occurrent preference model. What each person has most reason to do is whatever will best achieve his present or occurrent aims. It is instrumental in the sense that the aims are treated as given. Any aim can provide good reasons for acting and no aim can be claimed to be irrational. The deliberative model parallels Gauthier's considered preferences model. Some aims may not provide good reasons for acting. To judge whether this is so, an agent is required to consider the aims he would now have if he knew the relevant facts and was thinking clearly. Parfit's model is neither of these. He opts for the critical model. According to it,

> Some kinds of aims are intrinsically irrational, and cannot provide good reasons for acting whether or not they passed the deliberative test. What each person has most reason to do is whatever will best achieve those of his present aims that are not irrational.[148]

Not only are some desires intrinsically irrational, according to Parfit, other desires are rationally *required*.[149] "It is irrational to desire something that is worth *not*

---

[148] Parfit, 94.

[149] Parfit, 120.

desiring -- worth avoiding."[150] There is no way this position can be maintained without abandoning the reliance on individual preferences. The basis for such a claim can only be a standard-bound one; to make judgments on an individual's preferences from without, rather than from within. Adopting the critical present-aim theory is to abandon entirely a present-aim theory, or so it seems to me. Parfit solves one problem by reintroducing another. It is for this reason I have called my theory something different than Parfit's. My occurrent preference model could be instead called, following Parfit, the Instrumental Version of the Present Aim theory. This would not be a good advertising move, since Parfit is critical of that model. Let us briefly see why he is, and whether he is right.

Parfit introduces a variety of intuition-pumps to help support the need for a critical present aim theory. Imagine being future Tuesday indifferent. You are indifferent, we are asked to believe, about whatever happens on Tuesday. If we consider an event identical in every respect except that it falls now on a Tuesday, and now on a Wednesday, this future Tuesday indifferent individual will regard these otherwise identical events differently. If it is a pleasurable event, and he has a choice, he would prefer it to occur on Wednesday so that he could enjoy it. If it is a painful event, like an operation, he would prefer it to occur on a Tuesday, since

---

[150] Parfit, 122.

indifference is better than pain. What if, Parfit, asks us, he has a choice between a painful operation on Wednesday and a much more extremely painful operation on Tuesday, where the results of the two operations are identical. Our Tuesday indifferent man will prefer the more painful operation on Tuesday, and this Parfit reasons is extremely irrational.[151] Is he right? I don't think so. However odd, it is not irrational on the instrumental version. In fact, it is quite rational. If the man is indifferent to anything that happens on Tuesday, which we're led to believe, then it cannot be the case that the pain, measured phenomenologically at least, is greater on Tuesday than on Wednesday. Even if the pain would be far greater to anyone else, given his Tuesday indifference, it can not be painful to him. If I am indifferent to the pain on Tuesday, but not indifferent to the pain on Wednesday, then I too would choose the operation on Tuesday over Wednesday. It is only peculiar if I were not in fact Tuesday indifferent.

Parfit did not see it this way because he ruled out both the hedonistic and instrumental interpretations of self-interest. On most objective criteria, the Tuesday indifferent man is irrational. But I have defended the hedonistic version against Parfit's complaint, and moreover, the hedonistic version of self-interest is more consistent with contractarian strains of thought. Keeping a hedonistic view of self-

[151] Parfit, 124.

interest, we have no grounds to adopt Parfit's critical version of the present aim theory. In fact, because our intuitions are different depending on whether we have ruled out the instrumental-hedonistic version or not, using this thought experiment to support throwing out the hedonistic-instrumental model is simply an exercise in question-begging.

Perhaps the Tuesday indifferent case is too peculiar to be of much use. Parfit used it as a foot in the door technique. His belief was that we would all see such preference ordering to be utterly irrational. Once we accepted that, then he could show that in reality we have certain similar beliefs. For example, we have a tendency to prefer pains, otherwise avoidable, in the future to lesser pains in the present. To claim these sorts of things are irrational, as Parfit evidently wants, is to bring back the Prohibition era. Hangovers are unpleasant, and could have easily been avoided by abstinence. Unwanted pregnancies, overeating, not saying no to one's boss, all would become irrational acts and thus susceptible to outlaw. A more realistic theory, sensitive to psychological realism, will not be so strict with preferences.

We can now see that the occurrent preference theory is different than Parfit's present aim model in two respects. The occurrent preference model is instrumental and hedonistic (or phenomenologically dependent). The critical present-aim model is neither.

5.4 SUMMARY

It is rational to follow what is in our interests, and if this is best done by curtailing certain of our other interests, so be it. Curtailing one's interests in order to fulfil other interests one has is not a criticism or refutation of the self-interest theory, since the bottom line remains the satisfaction of whatever is in your interests; ie., what we most want or value. To achieve this, we do need a distinction between greater and lesser wants. But this cannot be decided on grounds external to the agent. What counts as lesser and what counts as greater must still be a decision of the agent herself; not some critical theorist on the outside looking in. One's occurrent preferences are whatever preferences one has come to embrace.

One thing I have not done yet is defend what is becoming more and more an antiquated theory of rationality: the instrumental model. I have discussed various authors' proposals for an alternative theory in chapter four. That these fail in one respect or another does not all by itself justify my maintaining an instrumental version. In the following chapter, I examine and defend against criticisms of the instrumental model.

Chapter 6

INSTRUMENTAL RATIONALITY

6.1 THE INSTRUMENTAL MODEL OF RATIONALITY

The occurrent preference model follows from adherence to the standard model of rationality; i.e., the instrumental model. This standard model has been criticized, however. A purely instrumental model of rationality is not going to back, at least easily, reasons to be moral. On the standard model of rationality, which I endorse, rationality, strictly speaking, can not care less whether one annihilates the human race so long as one fails to prick one's finger. To be moral, on the other hand, an individual must be more far-sighted than the strict instrumental model requires.

If what matters is what *would* in fact survive adequate reflection, then we are asked to imagine hypothetical situations rather than the real -- and this has the danger of allowing paternalism when we link our model of rationality to morality. It is *illegitimate* for a state to speculate what an individual *would* have preferred *if* he had considered it more. That is a task we can do for ourselves, but *not* for someone else. I might here specify *strangers*, especially. For our friends, those about whom

we know a fair amount, it may appear less obvious that we can not speculate about what *would* be beneficial to them. We may believe ourselves to be doing a friend a favour by interfering in one of his actions on the basis that he, in this case, did not fully consider the repercussions against his *other* preferences by following this particular one. Nevertheless, our friend is the final judge on whether we helped him. And this is the point I wish to emphasize: the sovereignty of subjectivity -- subjectivity that Gauthier himself endorses.[152]

Speculating on our own hypothetical preferences will not make them any more considered, however. Preferences that are classified as "considered" according to personal speculation simply equates "considered preferences" with those upon which there is no further personal deliberation. This will admit such a wide range of preferences that speaking of "considered" preferences becomes unintelligible. My claim in the preceding chapter was that if preference ordering is strictly subjective, then a "considered" preference is simply the occurrent preference that survived whatever amount of reflection one happened to have given it (which, note, may be no reflection whatsoever).

If the purpose of rationality is to further one's subjective preferences, it ought not matter what those preferences are (whether considered or not). If on a whim

---

[152] David Gauthier, *Morals by Agreement*, Oxford: Oxford University Press, 1986, ch. 2, sec.4.

Malcolm Lowry's ex-consul prefers to ride the ferris wheel, then he was rational (although drunk) to buy a ticket and strap himself in -- even if, moments later, he no longer *had* that preference and wondered if he ever had it. That the *content* of one's preferences are left unspecified is how the economists envisaged rationality. To allow the content of rational choice to be restricted, as Gauthier, Baier, Schmidtz, and Taylor want, is to be open to the charge that one's concept of rationality introduces *a priori* normative assumptions.

One possible way to rescue the notion of considered preferences is to argue that the criterion for distinguishing a considered from an unconsidered preference is to wait and see whether the occurrent preference was in fact satisfied by the decided action (or preference choice followed by action to satisfy *that* preference choice). If not, then the preference was not, after all, considered. There is some indication that Gauthier was thinking along these lines. He admits that "some processes of preference formation may eliminate any self-critical capacity" and that, as a consequence, "they would be ruled out but ruled out not in themselves *but for their effects*."[153] But this is really the criterion of rationality *simpliciter*. *Any* (considered or unconsidered) preference will do: if they are satisfied; the action was rational. If not, then the action was, after all, irrational. If the preference *was*

---

[153] David Gauthier, "Morality, Rational Choice, Semantic Representation", *Social Philosophy and Policy*, 5, 1988, 194, my emphasis.

satisfied, but led to *other* preferences being unfulfilled, we may again say that the action was irrational so long as the unfulfilled preferences were preferred over the fulfilled preference even at the time of choice. If two preferences are negatively related (e.g., the preference for smoking and the preference for health), and the second preference is *more* preferred, then it turns out to be *irrational* to have satisfied the first. This follows *without* appeal to considered preferences. Plainly, if preference ordering is strictly subjective, then a considered preference is simply the occurrent preference.

Clearly, to conceive of rationality as furthering one's preferences, some weighing of one's preferences must be done that is determined not by an antecedent criterion but by the agent herself. Once a given preference is weighted above the others, it is rational to maximize that one. To retain a concept of rationality that is a mere preference maximizing tool, and hence contentless, we can not determine in advance what preference *ought* to be given more weight. We can not even decide that the preference reflected upon the most ought to be given more weight. As Baier noted, some things will not change no matter how much more reflection one makes upon it.[154] The reflection required in determining that 5 + 7 = 12 is minuscule compared to the reflection required in assessing whether 37,698

---

[154] Kurt Baier, "Rationality, Value, and Preference", *Social Philosophy and Policy*, 5, 2, 1988, 42.

x 43,621 = 12,375,855,018. That the latter is therefore more "considered" is absurd. Certain Eastern philosophies, moreover, urge spontaneity as the surer guide to one's true preferences.

It appears, then, that we can not get any closer an approximation of the criteria for rationality than the standard instrumental model. All that we have is that if an act does not further the preferences one had intended to further, and no external countervailing circumstances are to blame, then there was some irrationality present. As Gauthier recognized, we can not be any more specific given that what is rational is at least some fit between an act and an intention. We can not say, "the *act* was irrational," for example, without adding, "given the preference intended to be satisfied." Alternatively, we could say, "the *preference* itself was irrational, given the act committed." At least in one circumstance this alternative view will not be wrong. Preferences must be *able* to be furthered if what is rational is to further one's preferences. But this is not at all what Gauthier means by "considered," and nor is it what Schmidtz had in mind by introducing maeutic ends.[155] The restriction is indeed very weak, for the criterion in deciding whether a given preference is rational must still be subjective. If one believes an act, x, *will* satisfy one's intended preference, one is certainly motivated -- and rational -- to

[155] See my discussion of Schmidtz in chapter 4, section 3.

enact x given one's beliefs.[156] And if after the act (which may fail to satisfy the actor's preference according to outside observers), the actor nevertheless *believes* that his (unrealizable) preference *was* satisfied, then the preference (and the act) was rational given that an act could satisfy it according to his (albeit errant) estimations. Thus the weak criterion that a preference must be realizable is itself limited to solely the *actor's* estimations. Otherwise, we would call those great athletes, artists, explorers, and scientists irrational for trying to accomplish that which others deemed impossible. The point, here, simply, is that preferences are not the sole factor of rationality. Rationality concerns the relationship between a preference and an act. And this is so since it is rational to further one's preferences.

6.2 INFORMATION

I claimed above that rationality concerns the relationship between preferences and acts. Acts must be conducive to furthering the given preference and the preference must be such that one can realize it. There is concern, however, that this is not a sufficient picture of rationality. Due to lack of information, or misinformation, perfectly reasonable acts may fail to satisfy what Baier would consider perfectly

[156] As Gauthier admits in *Morals by Agreement*, 30.

reasonable preferences. For this reason, some find it difficult to assess the rationality of acts without accounting for the availability of correct information. It would appear that information also plays a role.

Let us assume that under normal conditions, C_1, it would be rational for Betty to cross the bridge to satisfy her desire to reach the opposite side of the river. In condition C_2, unbeknownst to Betty, the bridge is out. Crossing the bridge in C_2 yields a result contrary to her interests: she falls into the river. Both the act and the preference are identical under C_1 and C_2, yet the result is disastrous in C_2, not in C_1. If we are to assume that Betty's crossing the bridge in C_2 was irrational, since it failed to satisfy her preference, then we must introduce information as a third condition of rationality.

This example exhibits a common fact of life; what satisfies a preference one day will not necessarily satisfy the same preference another day. Actions have to be assessed, and this must be done relative to available information. Introducing information as a condition of preference satisfaction, however, looks disastrous for a moral theory based on occurrent preferences. The moral emphasis on occurrent preferences depends upon a purely means-end instrumental model of rationality. What is rational for an agent to pursue is whatever best satisfies that agent's naked preference. I have throughout Part II of this thesis resisted the temptation to interpret what we mean by naked preferences as those preferences individuals

ought to hold. Yet, introducing the criterion of information on rationality does just that; at least if there is any externalist stricture on what counts as correct or adequate information. If so, we cannot hope to ground a rationalist ethic on occurrent preferences alone.

If morality can be enforced through social sanction, and what is moral is contingent on what is rational from the individual agent's point of view, and we discover that what is rational for that individual may well be not what he is aware of, given misinformation or lack of information, then we may legitimately intercede on that agent's behalf, even if that agent resists our interference. By the introduction of the concept of information in our notion of rationality, we have quickly moved from a fundamentally non-interfering philosophy to paternalism.

There is no need for paternalism, however. The role of information comes in at two distinct levels of discourse: rationality and morality. It is important to keep these separated. At the moral level, information does play a role. If Betty has paid a toll to use the bridge to reach the other side, she should expect the bridge will actually reach the other side. Her expectation arises from the implied agreement to pay to reach the other side. She paid, therefore she expects and is in fact entitled to reach the other side. If she is prevented from receiving her expectations in this case, she is entitled to receive compensation. One cannot be compensated without being informed of the need for compensation. She is entitled to information that the

bridge is out. Although she does not expect a cluttered list of things not expected to happen to the bridge during her crossing, she does expect to be informed of anything that is expected *out of the ordinary*. It is out of the ordinary for a bridge to be out. Hence, Betty has a right to expect a sign, at least, telling her of this extraordinary fact. The toll operator has a duty to inform Betty that the bridge is out.

If the bridge is publicly owned; that is to say every citizen has paid taxes to erect that bridge, the state in general has the obligation to inform Betty, and other tax-payers, that the bridge is out. This does not necessarily mean, however, that any passerby has a duty to tell her the bridge is out. Admittedly, this complicates matters. It may seem that although society in general has a duty to inform Betty, no one in particular has this duty. And this is tantamount to predicting that no one will in fact tell Betty, unless human nature is such that a certain number of us often act in the interests of others without being coerced. This may well be true. No moral theory will forbid anyone from imparting information on willing listeners. Moral theories differ, however, on whether there is a duty of a particular passerby to tell Betty that the bridge is out. The principle that "the parties involved are responsible" is problematic for democratic societies in which the claim is that everyone is involved. The more involvement, the more diffuse is the responsibility. This speaks toward privatization, but this is not my concern at this point. My main concern in this

section is not with the *moral* implications of information. Rather, I am interested to know whether information should be factored into our concept of *rationality*.

Information at the level of rationality is a different matter. According to one interpretation of the standard instrumental means-end model, what is rational to do is whatever will satisfy one's preferences. The converse of this is that whatever does not satisfy one's preferences is irrational. Taken strictly, then, Betty's crossing the bridge in C_2 is irrational, since it did not satisfy her preference. This is an absurd result (as I shall demonstrate below) and can be avoided quite simply if we take the instrumental model of rationality to claim that what is rational to do is whatever will satisfy one's preferences *under normal circumstances*. Under normal circumstances, it is rational to cross the bridge. Notwithstanding the disastrous results occurring from crossing the bridge in C_2, the act is no less rational for that.

How is the result absurd if we fail to interpret the instrumental preference-satisfaction model of rationality without this implicit appeal to normalcy? Ignoring the normalcy clause is to adopt a view of rationality that claims only acts which *successfully* achieve preferences are rational. Parfit defends a "success" version of rationality,[157] but I have already argued against it on the grounds that it appeals to standards separate from the occurrent preferences of the agent, and thus

[157] Derek Parfit, *Reasons and Persons*, Oxford, Clarendon Press, 1987, 149-150, 494-495.

abandons the neutral means-end model.[158] Another problem with it is that it is simply too restrictive. Many countervailing factors may intervene in the fulfillment of one's desires, and some of these have nothing to do with either the aim or the means under normal circumstances. If a bomb were placed under the bridge and exploded while I was crossing it, the failure of satisfying my goal can not be because I was irrational. If I go out on my lawn to relax and am hit by a drunk driver off course, I cannot be claimed to be irrational merely because I did not achieve a relaxed state. Under normal conditions, my actions would have achieved my aims.

These last two examples, more than showing the absurdity of success theories of rationality, indicate that we require a view of rationality that does not demand strict requirements about information. Richard Brandt, for example, declares that an act is rational if and only if it is accompanied by full information.[159] We do desire moral culpability for failing to provide information in certain cases. If every accident was your fault for being irrational (for failing to recognize the relevant information that indicated the accident was imminent), then no one can be culpable for harms. Had you known that I was going to shoot you today, you would not have come. Since you came, your coming was irrational. Since we are not

---

[158] See earlier in this thesis, chapter 5, section 2, "Hedonism and Death."

[159] R.B. Brandt, *Ethical Theory*, Englewood Cliffs, NJ: 1959, 173-174.

responsible for irrational acts of our neighbours, I cannot be responsible for shooting you. Clearly, this is preposterous.

Arguments that are unsound are not necessarily invalid and we may be justified in holding a belief even if the belief turns out to be false. In logic and epistemology, then, we do not hold success theories of validity or of justification. Neither should we hold success theories of rationality. Demanding that the criteria of rationality includes correct information makes the theory of rationality a success theory. This is so because any failure of a plan has a cause. Any cause can be understood in terms of information. Had this information been available to the subject, she would not have enacted that plan. Rationality must be less restrictive than that. A rational act appropriately uses the normal information at hand; no more is required. There can be no appeal to types of information that one should possess (at the level of rationality).

Information must play some role, however. What if Betty simply ignored the sign clearly posted that the bridge was out? We want to say, I presume, that Betty is irrational for ignoring that information since it has direct relevance to the satisfaction of her preference. We want to limit the criterion of information in regards to rationality to only that which is normal to the situation at hand, and no more. What counts as "normal," however, is difficult to define in general terms. The normalcy condition of relevant information permits acts that are accidentally

unsuccessful to be nevertheless deemed rational. That information plays some role secures the occurrent preference model against the charge that they hold every act to be rational simply for its having been acted. Rationality is properly open to public scrutiny. What I have argued for in the preceding chapters is that one's preferences are not open to public scrutiny. One's acts are, however. And one way acts go astray is through faulty information. The relevant information must be acquired or maintained by the agent in an epistemically responsible manner. What counts as epistemically responsible is also, unfortunately, open to debate.

The role information plays in theories of rationality is akin to the role the world plays. It is rational to satisfy my preference. To do this, I need to take some stock on the available means of satisfying my preference. As Gauthier recognized, part of this includes anticipating the preferences of others. Another part of the means to satisfying my preferences includes inspecting the available information. A sign claiming: "Bridge out!" is one of the pieces of information the subject utilizes in determining how best to get to the other side. She ignores this at her peril, given her own preferences. Rationality utilizes the available information. To ignore relevant information that is present is counter to rationality. This, however, is not a stringent requirement. There is nothing inherent in the concept of rationality, nor could there ever be, that determines in advance according to some standard what sorts of information is *required* for every potential act. Rationality is as neutral to the

content of information gathered as it is to the aims of the individual. Merely, albeit importantly, there must be some strictures on *how* that information has been gathered. It is not a good sign if it comes about through hasty generalizations, or messages from one's dead mother in dreams. Actions are cognitively accessible to epistemic criticism precisely in regards to how the information is gathered or maintained.

On the other hand, if rationality requires epistemically approved information, in what sense is this requirement not stringent? Even if the role information plays in rationality is delegated solely to whatever the individual singles out under normal conditions, there may be good reason to take stock of more information. The more information one has, up to a point of saturation, the more likely it is he will satisfy his preferences. Rationality's goal, recall, is to further preferences. If it is the case that gathering more information serves rationality's purpose, why then, an objector may persist, does rationality not require it?

There is a practical problem. How far afield does one search? Searching for information is understood as instrumental to achieving one's preferences (unless one prefers to search for information for its own sake). The longer one searches, the longer one must wait to satisfy the original purpose. Searching for information without limits will be counterproductive to the goal of satisfying one's preferences. How can we rationally put a limit on information searches? What would the general

rule look like? There are three options: (i) antecedent time constraints; (ii) impositions of quantity of information; and (iii) degree of conviction.

(i) How much time should we devote to information gathering? It will depend on how damaging preference-unfulfillment is, how impending the choice is, and how ambiguous the existing information is. How are these determined? By observing the external world. But to what extent are we required by rationality to make these initial observations? Asking this, I hope the reader sees, is peculiar. Surely these things are just given in context of the choice situation. If these are parts of the concerns that constitute the choice context, they are not themselves requirements of rationality, but the things with which rationality works; just like the preferences. As rationality cannot determine what preferences to have or not to have (other than showing which can and which cannot be realized in the world) rationality also cannot determine the urgency of the felt need of preference satisfaction, nor the seriousness the preference has for the individual. Nor can rationality determine that choices must be made under situations of clarity, since we often do not have the luxury of acting with clarity. If this is so, then time factors are a variable dependent on the given preference and the existing context in which it is embedded. It cannot then be a general criterion of rationality.

(ii) For information to be a condition of rationality, it must be a condition that is general enough to apply in all choice situations. Claiming that acts are rational

only so long as they have gathered x number of information bits is arbitrary. Such an imposition cannot successfully be applied across the board, unless the number is low enough to permit acting immediately on the following sort of algorithm:

preference = {want ice cream}
information = {see ice cream}
action = {eat ice cream}.

If this thought matrix has considered the requisite amount of information, then the level of information needed to be gathered is satisfied by all acts except for blind stumblings. Moreover, imposing a criterion of information on our concept of rationality in this manner ignores *quality* of information. One good piece of information may surely suffice, especially over five more bits of false, irrelevant, or redundant information.

(iii) The last possible rule for information gathering says something like: gather information until you are satisfied. Typically this is true. We gather more information only if we are not satisfied with the information that we presently have. And we stop gathering information when we think we have enough under the circumstances. Placing the emphasis on personal satisfaction conceives the concept of rationality in a hedonist manner. The deciding vote is the rational agent herself. This has been precisely my point all along. If the agent herself determines the appropriateness of her information, then no external standard of appropriate

information can have weight over the individual. To recognize this is to concede that correct information is not a necessary criterion of rationality.

Let us return to Betty's situation. Should she cross the bridge? Is there any other information that is relevant? In fact, yes, but *ex hypothesi* it is unknown to her. Hence, the need for further information is not seen as needed by her. Betty is still rational for crossing the bridge, given her aims and given the information she had. If either she gets new information, or alters her preference, then we must reconsider the rationality of her continuing across the bridge. It is to her advantage, we might say, for her to have the added information, but this does not detract from the rationality of her action; we are assessing her action *from her own standpoint*, and not from someone else's. It is rational to assess the information *one has*, but no stipulation can be made at the level of bare rationality to seek out more information. What is irrational is to act contrary to one's occurrent preferences and given information.

The problem with incorporating information into the model of rationality, even to the limited extent that I have tried here under conditions of normalcy, is that it is not clear that many of us can meet the requirements most of the time. This would be tragic for an occurrent preference model. A moral theory that pretends to appeal to the preferences individuals actually have will fail to convince if it demands epistemic responsibility beyond psychologically realistic boundaries. If what it takes

to be epistemically responsible is more than many of us are in fact, it fails to be a motivating model. Nevertheless, we need a model of rationality that permits criticism of the connections between means and ends. Information plays some role in this, but it is a limited role, restricted to what is available given the immediate context.

6.3 INTRANSITIVITY AND INDIFFERENCE

The instrumental model of rationality has recently been under attack from another quarter. The instrumental model of rationality has it that an act may be rational so long as it furthers one's preferences. According to von Neumann and Morgenstern, however, those preferences must themselves meet certain conditions.[160] These constraining conditions are completeness, transitivity, reduction of compound lotteries, continuity, substitutability or independence, and monotonicity.[161] Most of these assumptions or axioms have met with some complaint.[162] For example, many now believe that intransitive preferences are far more common and reasonable than

---

[160] John Von Neumann and Oskar Morgenstern, *Theory of Games and Economic Behaviour*, Princeton: Princeton University Press, 1944.

[161] See R. Duncan Luce and Howard Raiffa, *Games and Decisions: Introduction and Critical Survey*, New York: Dover, 1957/1985, 23-31.

[162] Notably from Daniel Kahneman and Amos Tversky. See, for example, "Rational Choice and the Framing of Decisions," in Karen Cook and Margaret Levi (eds) *The Limits of Rationality*, Chicago: University of Chicago Press, 1990, 60-89.

previously suspected, and that therefore no proper theory of rationality should exclude them.[163] The implications of a view of rationality that embraces intransitive preferences have not yet been fully catalogued, but I wish to point out here that permitting rather than forbidding intransitive preferences fits rather than clashes with the occurrent preference model.

It has generally been thought that so long as one's action can reasonably be expected to satisfy one's preferences, that action is rational. On this model, there is no constraint on what those preferences are, so long as they can in general be satisfied. Preferences that cannot reasonably be satisfied are alone deemed unfit candidates. Typically, this has been thought to be the case with intransitive preferences. If one's preferences are intransitive, then no action can satisfy one's preferences, since necessarily, another action or choice would always be preferable. If Hector prefers A over B and B over C, but C over A, he will have no rational choice available. If he chooses B, he is worse off than had he chosen A. Had he chosen C, he is worse off than had he chosen B. But if he had chosen A,

---

[163] George Schumm, "Transitivity, Preference, and Indifference," *Philosophical Studies*, 52, 1987, 435-437. Howard Sobel, *Taking Chances: Essays on Rational Choice*, Cambridge University Press, 1994. John Broome, *Weighing Goods: Equality, Uncertainty and Time*, Oxford: Basil Blackwell, 1991. Graham Loomes and Robert Sugden, "Regret Theory: An Alternative Theory of Rational choice under Uncertainty," *Economic Journal*, 92, 805-824. Jean Hampton, "The Failure of Expected-Utility Theory as a Theory of Reason," *Economics and Philosophy*, 10, 1994, 195-242.

he is worse off than had he chosen C. It is no wonder, then, that so long as Hector's preferences are intransitive in this way, he is considered (or until recently would have been considered) irrational for having such ill-ordered preferences. There is a constraint, then, of sorts, on the preferences an individual may have to be counted rational. An individual's preferences must be well-ordered on the traditional instrumental model of rationality.

Schumm shows how otherwise normal and reasonable preferences, intermixed with indifferences, can create intransitivity. Imagine we have three boxes to choose from. In each box there are three marbles (or Christmas ornaments, or socks, or what-have-you) of three colours each: red, blue, and green. The marbles are identical, but the hues are slightly different. Box 1 contains r1, b1, and g1, Box 2 has r2, b2, and g2 and Box 3 contains r3, b3, and g3. Our supposition is such that of the three red marbles, you prefer r1 over r2, but are indifferent between r1 and r3 as well as between r2 and r3. Since r1 is in Box 1, and r2 is in Box 2, then everything else being equal, you would prefer Box 1. Everything else is not equal, however. Of the green marbles, you prefer g2 over g3, but are indifferent to the subtle hues of g1 over g2 or g3 over g1. Thus, looking only at the desirability of the green marbles, you would prefer Box 2 over Box 3, since you prefer the green hue of the marble in Box 2 over the green hue of the marble in Box 3. Meanwhile, you happen to prefer the blue colour of the marble in the third box over that of the blue

marble in the first box (b3 over b1), but cannot tell b1 from b2 or b2 from b3. Given these preferences and indifferences, one can see from the chart below that there is no box that will be most preferred.

Box1		Box2		Box3		Box1
r1	>	r2	=	r3	=	r1
b1	=	b2	=	b3	>	b1
g1	=	g2	>	g3	=	g1[164]

If you are given a choice between only two boxes, your preferences can yield a choice. You would prefer Box 1 over Box 2, because you prefer r1 to r2 and are indifferent to the differences between g1 and g2 as well as between b1 and b2. If, on the other hand, you were given a choice between Box 2 and Box 3, you would prefer Box 2 because although you are indifferent between the reds and the blues in the two boxes, you prefer the green colour in Box 2 over the green colour in Box 3. If, instead, you were given only the choice between boxes 1 and 3, you would pick Box 3 because you prefer g3 over g1 and for all intents and purposes you cannot tell the other colours apart. What this particular preference ordering means, however, is that although you will prefer Box 1 to Box 2, and Box 2 to Box 3, you will prefer Box 3 to Box 1. This proves that your otherwise reasonable preferences are intransitive and thus incapable of yielding rational choice.

---

[164] Read "r1 > r2" as "r1 is preferred over r2"; and "r2 = r3" reads "r2 is for all intents and purposes indistinguishable from r3."

What Sobel and Schumm have appeared to show, then, is that intransitive preferences, that can obviously yield no rational choice, is a result of occurrent preferences. So long as preferences are unconstrained, it is not the case that rational choice can follow in all cases. The question I wish to address concerns the implications of permitting intransitivity in our model of instrumental rationality.

Previous attempts at preserving the standard consequentialist models of rationality against this (or similar) problem(s) have been somewhat *ad hoc*. If part of the definition of rational preferences is that they be action-guiding, then however reasonable the origins of such intransitive preferences, it is not rational to maintain those preferences. Hansson, for example, adopts this approach.[165] Gauthier, too, in solving prisoner's dilemmas opts for altering one's preference schemas on the basis that the old preferences are unproductive (if not intransitive).[166] David Schmidtz argues that a proper model of rationality can assess the preferences or ends one has on the basis of higher-order, or *maieutic* ends.[167] However reasonable

---

[165] Sven Hansson, "Money-Pumps, Self-Torturers and the Demons of Real Life," *Australasian Journal of Philosophy*, 71, 1993, 484.

[166] Gauthier argues that it is rational to alter our dispositions from being a straightforward maximizer to becoming a constrained maximizer (Gauthier, *Morals by Agreement*, 167-189).

[167] David Schmidtz, "Choosing Ends," ch. 3 in *Rational Choice and Moral Agency*, unpublished manuscript, 1994.

these approaches appear, they put the cart before the horse. We now no longer define rational acts according to preferences held, but determine what counts as rational preferences according to a preconceived outcome -- an outcome independent of one's occurrent preferences; preferences that may in fact be intransitive. Rationality on such a view loses its tie with preferences and becomes, instead, standard bound.[168] To preserve the concept of rationality as a tool to satisfy (or maximize) preferences while at the same time admitting that many common preferences are ill-ordered is to concede that rationality cannot be action-guiding in all circumstances.

Nothing should be surprising in this admission. Moreover, there should be nothing damning against a preference-based model of rationality that concedes that rationality cannot be action-guiding in all circumstances. The commonality of intransitivity of preferences stems, we have seen, from indifference. That we are indifferent between two options makes it impossible to discover a uniquely rational choice between them. That is not unusual. There is not any rational choice to make between them on the simple grounds that we have no preference to satisfy between these two options. The instrumental model of rationality is not stymied because of

---

[168] For a survey of the dangers of standard-bound concepts of rationality, see Anthony de Jasay, *Social Contract, Free Ride: A Study of the Public Goods Problem*, Oxford: Oxford University Press, 1990, 100-101.

the existence of indifferences. It may yet be rational to choose one of them, rather than none of them. A simple reminder is the tale of Buriden's Ass. Being indifferent between the two haystacks, the ass cannot make a rational choice. Failing this, it remains standing in the middle like a computer program in a loop, unable to execute any further command. The result is it starves, having failed to choose either. If the sole preference of the ass was to eat from the best haystack, and that both haystacks were equal, then its preference, as put, could not be rationally satisfied. It is highly unlikely, however, that eating from the best haystack is the only operative preference the ass has. A higher-order preference, operating simultaneously with the preference to eat from the best haystack and in fact fuelling that preference, is the preference to simply eat. Given this higher order preference of the ass, the ass's indecision is clearly irrational.

There is a difference, of course, between indifference and intransitivity of preferences. In cases of indifference we admit that no choice will ever be the most preferred choice, and acting on one rather than another does not mean we will inevitably be *disappointed* by that choice. This is not the case with intransitive preferences. Acting on one of a set of intransitive preferences means that we will necessarily be disappointed. Since we will be necessarily disappointed by any of our choices, however, we are indifferent to that set of choices.

The emphasis on occurrent preferences, recall, requires there be no external constraints on preferences. So long as one's actions satisfy one's preferences, that action is rational. And since moral acts are a subset of rational acts, at least within the narrow ethics tradition, then the same applies to moral acts; that they satisfy individual preferences. If one's preferences are intransitive, however, no action can satisfy one's preferences; another action or choice would always be preferable. As we have seen, making a censure against intransitive preference orderings would be arbitrary and *post hoc*. The desire to avoid such imposed strictures against individual preferences fits with the occurrent preference model. The occurrent preference model wishes to place no constraint on individual preferences, and nor need they. So long as one's preference orderings are intransitive, that individual will be unable to effectively pursue her preferences. The constraint on rationality is her own; it is not imposed on her from theoreticians. It is simply the imposition of constraints on preferences from theory, rather than the subject, that is disapproved of by the occurrent preference model. No pretensions are made that everyone have perfectly consistent preference orderings.

PART III

OCCURRENT CONTRACTARIANISM

Chapter 7

OCCURRENT CONTRACTARIANISM

In chapter four, a distinction was made between preferences that were considered, and preferences that were not. The argument there supported abandoning the recent move in contractarianism to base ethics on considered preferences. In chapter five we introduced the theory of occurrent preferences. Now we are in a better position to see more fully why reliance on considered preferences becomes no longer a preference-based theory. Reliance on considered preferences as a basis for moral theory introduces standards by which we evaluate individual preferences. Reliance on standards despairs of the original goal to link ethics to individual preferences; it does not embrace that goal.

Without reliance on considered preferences, it appears we cannot escape prisoner's dilemma situations. Yet reliance on considered preferences leaves the very origins and impetus of contractarianism; to ground morality on individual self-interest. Thus we are at a crossroads. Should contractarians abandon their original goal of linking ethics to individual preferences? Or can a wholly preference-based version of contractarianism succeed at grounding morality?

I shall argue for the latter. In this chapter, I propose a theory that is preference based without reliance on considered preferences and that is nevertheless sufficient to engender moral constraints on human action. I call it "occurrent preference-based contractarianism" to distinguish it from theories that place constraints on preferences. For short, we may refer to it as *occurrent contractarianism*.

7.1 OCCURRENT CONTRACTARIANISM

Self-interest must be determined by the individual alone. What the experts decide is in my interest may be of no interest to me. Basing morality on self-interest is the fundamental germ of contractarianism, and any reading that subverts this is to be resisted. Contractarianism must remain preference-based, rather than standard-bound, in order to remain consistent with the notion that morality should be motivating to rational self-interested people. Occurrent contractarianism is an open-ended preference-based ethic that defines legitimacy according to the occurrent preferences the concerned parties actually have. We do not care whether their preferences are considered to the requisite degree or not, so long as those are the preferences they actually have.

One's occurrent interests are whatever interests an individual happens to hold, somewhat consistently, over a period of time. They are not simply "current"

interests, for occurrent interests are a little more steadfast than current interests. But neither are they standard-based interests, for it is still questionable whether one should hold these interests occurrently given someone's estimation of reasonableness. Occurrent interests may well collide with standard-based interests. It may be in my occurrent interest to smoke, for example, but it is not likely to be in my interest according to a standard of health. If goodness is defined by one's occurrent interests, then it is good for me to smoke (ignoring the complication of second-hand smoke). Whereas, if we look to see what is good for me *in reason if not in passion*, then it is bad for me to smoke. There is danger of false stepping if we move from the one (grounded in human psychology) to the other (more morally appealing) without further ado.[169]

Moreover, and importantly, the criteria for consistency and duration is not a theoretical imposition about the sorts of preferences rational or reasonable individuals should possess. It is personal experience that determines the success of particular preferences. Preferences, let us not forget, are useful to individuals only so long as they can be satisfied. The satisfaction of preferences is not solely an individual matter. It depends on their social-psychological environment. The

[169] Recall (from my discussion in chapter 5, section 1) that Gauthier argues that a preference to smoke cannot rationally be pursued, given the information that smoking is bad (David Gauthier, *Morals by Agreement*, Oxford: Oxford University Press, 1986, 34).

individuals then, in this sense, do not choose the preferences that they hold somewhat consistently over a period of time. These preferences evolve, as it were, from the social, psychological, and historical factors in which the agents are embedded. Occurrent preferences are fashioned from the individual's abilities to satisfy initial preferences within the constraints of the real world. It is important to recognize that this constraint, although external, is not theory-driven. It is therefore not standard-based in the relevant sense. Occurrent preferences accrue from personal experience alone. The individual is still the sole judge.

I cannot say that Hobbes was either a standard-based contractarian or an occurrent contractarian. Reading Hobbes as a standard-based contractarian, although consistent with his premise that people would agree to a sovereign *in reason* if not in passion,[170] is not consistent with his premise that people define what is good based on their own interests.[171] We could make this consistent by claiming that what we mean by people's "own interests" are what they would agree to in

---

[170] Thomas Hobbes, *The Leviathan*, Buffalo: Prometheus Books, 1988 [1651], xiv (75); Thomas Hobbes, *De Cive*, Sterling P. Lamprecht (ed.) New York, 1949, 3.31 (57-58).

[171] Hobbes, *Leviathan*, vi, (24). "For there is no such *Finis ultimus* (utmost ayme), nor *Summum Bonum*, (greatest good,) as is spoken of in the Books of the old Morall Philosophers" (Hobbes, *Leviathan*, xi (49)). See also Hobbes, *De Cive*, 14.17 (166). For modern day accounts of the viability of this premise, see Owen Flanagan, *The Varieties of Moral Personality*, Cambridge Mass: Harvard University Press, 1991.

reason. But this does not capture the flavour of the psychological premises Hobbes endorses, for we are given the distinct expectation that an individual's notion of the good is sacrosanct no mater how perverse or unreasoned. The interests that defined individual good, at least in the Hobbesian state of nature, must be occurrent interests. To remain consistent with Hobbes's methodology of deriving an ethic from non-moral beginnings, we must, then, never supplant our occurrent interests with standard-bound ones.

7.2 HYPOTHETICAL CONTRACTARIANISM

Occurrent contractarianism places great emphasis on the preferences individuals actually have. Recent trends within the contractarian tradition focus on *hypothetical* preferences, preferences individuals *would* have if suitably placed. At first glance, this appears to be *opposed* to occurrent contractarianism for hypothetical preferences are, by definition, not the preferences individuals happen to have, but merely suppositions about what preferences they would have under different circumstances. This is mistaken. Satisfactory versions of hypothetical contractarianism must be dependent on occurrent preferences, not supplant them.

Recall that the acid test for morally permissible acts under hypothetical contractarianism is in deciding what would be *reasonable* for an individual to have consented to, suitably situated and with sufficient knowledge.[172] Two problems emerge for the hypothetical approach to contractarianism. One is that the agreement falls out of the picture entirely. Someone may explicitly agree to something, but had he been suitably placed and possessed sufficient knowledge he may not have made that explicit agreement. Buying a used car, for example, especially for the mechanically disinclined, is often a matter where further information would have affected the agreement. It is no longer clear, then, whether there are grounds to claim the bargainer is not liable to fulfilling his agreement. On the one hand, he made the agreement and thus is liable. On the other hand, had he been suitably placed, he would not have made such an agreement, and thus, *hypothetically speaking*, the contract is null and void. Under hypothetical contractarianism, moral grounding may have nothing whatsoever to do with the agreements in fact. For this reason, Harman believes Gauthier's book, *Morals by Agreement*, is a misnomer.[173] Hypothetical contractarianism, then, looks more like

---

[172] See my preliminary discussion of hypothetical contractarianism in chapter 1, section 3.

[173] Gilbert Harman, “Rationality in Agreement: A Commentary on Gauthier’s 'Morals by Agreement’,” *Social Philosophy & Policy: Gauthier’s New Social Contract*, 5, 2, 1988, 1.

an appeal to standard-based ethics since moral permissibility is no longer founded on the preferences individuals happen to have, but on those they *ought to* have, given suitable placement. The issue is not did they agree, but *should* they have. Whereas under preference-based versions of contractarianism the agreement is what is sacrosanct. So long as the people agree, we must take it that they, at any rate, believe their agreement is in their best interests, given their situation. Otherwise they would not have agreed to it.[174]

The other problem follows from removing the emphasis on agreements. Instead of agreements, the morality of actions must be based on what is rational, or reasonable, for the concern is what *rational* people, suitably placed, would reasonably agree to. Clearly these concepts have an important link to morality, but they cannot be a grounding for morality without the concept of agreement. Rawls, for example, considers that what is reasonable includes proposing fair terms of cooperation with others and being willing to abide by those terms.[175] It is true that reasonable and fair proposals offered by individuals willing to abide by the terms of the proposal will increase the likelihood of an agreement. And this is important

---

[174] With this, of course, comes a buyer's beware caveat. Some find this troubling enough to warrant forms of paternalism.

[175] John Rawls, *Political Liberalism,* New York: Columbia University Press, 1993, 49.

given individuals' desires for the mutual benefit accrued from agreements. Nevertheless, the proposal must be accepted, not merely proposed, and certainly not refused, if acting on it is to be considered moral. However reasonable an offer is, we cannot act on it if it does not meet with the agreement of all concerned parties. Agreement cannot be taken out of the picture.

Perhaps most problematic for the standard-based model that bases interests on what is rational is the plain fact that we are not, after all, very rational.[176] Given this, what people agree to *if they were rational*, amounts to saying what is moral is what people *will not likely* agree to in fact (being not rational). And *this* unmotivating notion of standard-based agreement is not worth the paper it's not written on.

Both problems are due to a misread, fortunately. Hypothetical contractarianism is applicable where we do not know of the reactions of others. There is no data on their occurrent preferences. Nevertheless we can imagine in certain cases how they would likely react, assuming a certain degree of normalacy. Occurrent contractarianism cannot be action guiding should the actor be considering the likely preferences of others for an act not yet committed. Here, hypothetical contractarianism appeals to the *likely* preferences rational individuals

[176] A. Tversky and D. Kahneman, "Judgment Under Uncertainty: Heuristics and Biases," *Sciences*, 185, 1124-1131.

would hold. Nothing that I have said about occurrent contractarianism will run counter to this notion of hypothetical contractarianism.

Likewise, the complaint that hypothetical contractarianism idealizes rationality beyond practical import is mistaken. Quite the opposite is the intent. Under hypothetical scenarios, we are concerned not with what is rational for people to prefer, *per se*, but what people will *likely* prefer. Since there is some degree of guess work in such vague considerations (recall that we are considering the reactions of others to acts we are merely *considering* committing), we assume the agents are rational -- not to the degree that their preferences are likely to be different than what they are under occurrent conditions, but simply to rule out a range of possible abnormal preferences. I cannot decide that it is morally permissible for me to burn down Betty's house because I assume Betty would like that. What I must assume is that she would not like that. If she really did like her house being burnt, that would be merely bizarre, and could not by itself justify my burning her house down if I had no antecedent reason to suspect that preference. We want to constrain hypothetical scenarios to reality or to real occurrent preferences as much as possible. The claim, "...what people would rationally prefer under suitable circumstances..." is not meant to rule as inadmissible the majority of human preferences. The complete opposite is its intent: to include real preferences

of the majority of real people. These normal run-of-the-mill occurrent preferences are precisely what contractarianism is asking us not to violate.

Not all hypothetical versions of contractarianism, I admit, is as strongly committed to psychological realism as Gauthier's model. Rawls's version, for example, is not. When Rawls speaks of people being "suitably placed" he means something quite different from Gauthier. All Gauthier means is a qualifier in the lines of "all other things being equal." If for some reason Betty thought I was suggesting to leave her alone forever, while I really meant to burn down her house, we can say she was not suitably placed to properly hear my request, and thus her affirmative response is not to be taken at face value. Rawls's notion of being suitably placed means precisely *not* the normal state of affairs. He means, what would people agree to, not under their normal situations, but under a *veil of ignorance*.[177] The "suitable place" in Rawls's case is a place no one has ever been in, could be in, nor, given their occurrent interests, would they ever *want* to be in. We should be wary that hypothetical contractarianism can come in two forms, realistic (or occurrent-preference sensitive) and unrealistic (or occurrent-preference insensitive).

---

[177] John Rawls, *A Theory of Justice*, Cambridge, Massachusetts: The Belknap Press of Harvard University Press, 1971.

7.3 SENSITIVE STANDARD-BASED CONTRACTARIANISM

In part two, we dismissed standard-based contractarianism in favour of preference-based contractarianism. Perhaps there are two types of standard-based contractarianism. One may be more sensitive to occurrent preferences than the other. The less-sensitive one may find grounds to do something to you against your occurrent wishes even if you never come to see the benefit of it. This insensitive version is simply standard-bound paternalism which I have already argued is antithetical to contractarianism proper. The more-sensitive one, however, may claim we or the state are permitted to do something to you against your occurrent wishes so long as you will, in time, come to appreciate it. Let us imagine that Scrooge will thank us later if we now abduct him and force him to work for Mother Theresa in Somalia.[178] So long as Scrooge will himself thank us in time, once he has undergone the appropriate change, then it is morally permissible to abduct him now on this more occurrent preference-sensitive standard-based contractarian grounds. Of course, if we abduct him now he will be screaming and kicking. His *occurrent* preferences we ignore.

Occurrent-time-sensitive standard-based contractarianism, however, violates the same basic contractarian tenet as the insensitive version. Both amount to

---

[178] This example comes from discussions with my colleague Alix Nalezinski.

denying the contractarian premise that people decide their own good, and that morality is simply a device to secure the free pursuit of those goods, no matter what they are. If this is the germ of contractarianism, even this more sensitive standard-based contractarian doctrine abandons contractarian principles for having the standards decide the good of the people rather than the people. Let me explain.

7.3.1 Drug Rehabilitation

Preferences are to some degree fluid, and are subject to change given the circumstances one is in. That I prefer Betty and not her identical twin Sue may be partly due to the happenstantial fact that I met Betty in a certain circumstance and Sue in a different circumstance. Circumstances influence preferences. This psychological fact lends support to the claim that it is perfectly legitimate to force one person into a situation knowing that through the process that person's preferences will come to change, and in time appreciate that coercion. Forced drug rehabilitation, for example, seems like the sort of thing that can be justified on this occurrent-sensitive but standard bound version of contractarianism, as with many therapeutic treatments. Importantly, forced drug rehabilitation and non-consensual therapy do not seem to be justifiable on straight occurrent contractarian grounds.

A dark side to occurrent-time-sensitive standard bound contractarianism is that this is the principle of all hypnosis or brain-washing techniques. We cannot

justify lobotomy merely on the grounds that after the operation the patient can be made to be happy with it. So the justifiability of what we may call "therapeutic" coercion must be grounded on a principle other than what the individual will feel about it after the event. This principle can only be standard-bound, not occurrent preference-based. Drug addicts typically feel horrible at certain times. During these times, they are quite capable of wishing to no longer be drug addicts. They may even voice this preference. That they nevertheless remain addicted is precisely part of being addicted. That we take the initiative to lock a drug addict in a room and forbid his taking more of the addicted drug may well be in line, then, with one of the addict's occurrent preferences. This is the reason for its justifiability, if it is justified. Therapy appeals directly to one of the patient's occurrent preferences, not to what their occurrent preference would be after the transformation.[179]

7.3.2 Surprise Parties

There are still some hard cases, however. A surprise party for Betty is an innocuous activity that appears to be justified, if justified at all, only by the welcome it would

---

[179] This is in perfect agreement with Frankfurt. "The predicament of the unwilling addict is that there is something which he really wants to do, but which he cannot do because of a force other than and superior to that of his own will" (Harry Frankfurt, "The Importance of What We Care About," in H. Frankfurt, *The Importance of What We Care About*, Cambridge: Cambridge University Press, 1988, 87).

receive after the fact.[180] Perhaps Betty has never voiced her desire for a surprise party. Giving her a surprise party, then, can not seemingly be justified in the occurrent lines as described above. With subtle imagination, we can see that surprise parties and non-consensual therapeutic procedures, if they are to be justified at all, can be justified on occurrent grounds solely. To give a surprise party, I am not acting on an unrealistic assumption of what Betty's response will be. Rather, knowing Betty the way I do, I can safely bet she would enjoy a surprise party. I make this educated supposition based on her occurrent preferences which I know. For example, she likes parties. That there is room for an occurrent justification can be more easily seen by trying to justify giving a surprise party to a stranger. If fifty percent of the population dislike surprise parties, I have a fifty percent chance of failure in pleasing the stranger. Going ahead with it, then, would be difficult to justify on occurrent grounds. If he is pleased after the fact, and thanks me for it, this is more by fluke than providing justification to my actions. Or consider giving a surprise party to someone we have good reason to suspect greatly *dislikes* such things. I submit that we would not be justified in doing so even if the recipient changed her mind after the event.

---

[180] See Jan Narveson (*The Libertarian Idea*, Philadelphia: Temple University Press, 1988, 182) for a brief discussion of this.

The difference in the antecedent justification between the three cases (giving a surprise party to someone who likes surprise parties, to someone we have no idea whether they like surprise parties, and to someone we have good reason to suspect greatly dislikes surprise parties) must be based on the occurrent preferences of the individuals we are dealing with. This should indicate that these cases are not to the point. They do not show where "considered" preferences trump occurrent ones, not in these otherwise innocuous scenarios, let alone in cases of much greater social significance.

7.3.3 Suicide

The above response can also be employed in the following sort of case: What if someone you know quite well intends to commit suicide, and you know they would rather not kill themselves if only they had more time to think about it under better situations. Does occurrent contractarianism forbid interfering in such cases? If so, so much the worse for occurrent contractarianism.

The answer is that stopping your friend from suicide does not (at least necessarily) abrogate occurrent grounds for contractarianism. If you know someone quite well, this means, we assume, you also know their occurrent preferences. Recall that occurrent preferences are not mere whims, although they can be. They tend to be longer standing dispositions and preferences. To know someone and to

know she is in a particular circumstance that would prevent her from recognizing her own occurrent preferences gives you reason to step in. So long as we act on occurrent preferences we know she endorses, this is consistent with occurrent contractarianism. You are acting on what you have good reason to assume to be her occurrent preference; what she would likely prefer given normal circumstances. Acting on realistic or occurrent-preference-sensitive hypothetical preferences does not violate the intent behind occurrent contractarianism. There would be a violation of occurrent contractarian thought if we act instead on merely attributed preferences; preferences she may not endorse, but which we have falsely attributed to her. Pseudo-psychological claims to unearth one's "true" occurrent preferences from beneath the veneer of suppression are to be suspected. It is not an occurrent preference unless it is actually endorsed by the agent herself.

7.3.4 Unconsciousness and Consent

What about operating on unconscious patients? Do doctors who operate without consent violate contractarian moral codes? In practice, unconsciousness is equated with consent.[181] We assume that the patient would consent to be treated if he were conscious. Being wrong about this does not mean the operation was not justified,

[181] For an informative discussion on this matter, see Gerald Dworkin, *The Theory and Practice of Autonomy*, Cambridge: Cambridge University Press, 1989, 115-120.

after all. We were justified, but wrong in this case. And this shows that justification of acts on others is grounded on occurrent preferences, and not on the preferences people have after the fact. For we do not say the operation in this case was unjustified merely because it failed, or had the unconscious patient known, he would have refused. The operation was justified based on assumed occurrent preferences, or hypothetically defined occurrent preferences. And we were justified in making that assumption given the occurrent preferences most people have for survival.

Showing the justification of drug rehabilitation, surprise parties, preventing suicide attempts, and operating on unconscious patients does not require an appeal to standards beyond normal or hypothetical occurrent preferences.

7.4 FURTHER CONSTRAINTS?

Occurrent contractarianism has other problems, mind you. The largest is its apparent inability to ground any duty to stick to one's agreements. That I borrow money today on the condition of repaying it with interest on Tuesday may well be in my occurrent interest today, but not necessarily in my occurrent interests on Tuesday. Basing morality on occurrent interests seems, in fact, to be the very antithesis of what we want. Occurrent interests are notoriously vain, base, ignorant, lazy, sensuous, and short-sighted. Morality, one would think, is designed to

overrule these sorts of petty motivations. Moreover, even granting that not all occurrent interests need be immoral, the very fact that I may have conflicting occurrent interests does not bode well for a theory that hopes to be action-guiding. Surely some constraint on occurrent interests is needed if we wish to base the institution of morality on them. That it appears in my self-interest to burn Betty's house down does not make it thereby morally permissible to do so!

Further constraints are typified by Gauthier's proposal of considered preferences. I believe, however, that further constraints on occurrent interests are unnecessary within the occurrent contractarian tradition. Contractarianism is not a form of ethical egoism. It does not say simply: "Do whatever is in your occurrent interests, however unconsidered." This is the state of nature, after all, and contractarians are anxious to escape the state of nature. The reason contractarianism is not a form of ethical egoism is because of its emphasis on agreements. As Gauthier argues, the moral sphere is not a *parametric* game, where actors take themselves to be the sole variable in a fixed environment. Rather, agents interact with others, and interaction involves *strategic* choice,

> in which the actor takes his behaviour to be but one variable among others, so that his choice must be responsive to his expectations of others' choices, while their choices are similarly responsive to their expectations.[182]

---

[182] Gauthier, 21.

What is in my interest must be tempered by necessity with what is in your interest. If we wish to interact, our individual interests must be compromised somewhat. My planned behaviour toward Betty is inadmissible, not because it is not in my occurrent interest; nor because it is not in my considered or standard-based interest, but simply because it is not (or not likely to be) in Betty's interest. Occurrent self-interest is already naturally constrained by the social-psychological conditions in which we find ourselves. Occurrent contractarianism emphasizes the furthering of self-interest (at least as far as morality is concerned) *within the confines of an interpersonal arena*.[183]

The contractarian credo is this: Whatever an individual has agreed to is morally permissible, so long as the agreement was voluntary and no one not party to the agreement was adversely affected. It is necessarily morally permissible since morality for contractarians concerns human interaction. That an agreement was made shows that all parties concerned agree. There is no more to morality than making sure this minimal condition is met in any interaction. Or at least, any further demands of morality become less motivating from that point forward. Occurrent self-interest comes into the picture as a matter of psychological realism. What I

---

[183] See earlier in this thesis, chapter 2, section 1, for a defence of this statement.

would agree to will be whatever improves my baseline by my own estimations. It is mere psychology to point out that this is the motive behind our actions. Where agreement is, there self-interested motivations are satisfied. Hence, the agreed upon arrangement, *whatever* it is, is necessarily rational. And so long as cooperation, as evidenced by making and abiding by agreements, is the cornerstone of morality, rationality and morality have a clear link.

Morality is tied to occurrent preferences, however, only so long as agreements are kept. But the rationality behind making agreements is not necessarily the same thing as the rationality behind keeping agreements. In the following chapter, I shall explain why it is rational that we would adopt an enforcement structure to ensure against defection, while at the same time explain why this does not abrogate the supremacy of occurrent preferences.

Chapter 8

HOW PREFERENCE-BASED CONTRACTS ARE BINDING

The genesis of contractarian theory is to ground ethics on individual preferences. Contractarianism is committed to being a preference-based moral theory. In order to serve the fundamental purpose of concerned parties, contractarianism must in some manner give rise to a choice-rule which is preference-dependent and standard-independent. The rule must produce the social choice from the orderings of the individual preferences. It must not produce it from anything else. Domain restriction contradicts this principle. It entails that preference does not prevail; rather some standard does. As de Jasay recognized, if a standard is decisive, it either produces what is preferred, in which case the standard is redundant, or it overrides preferences and produces a standard bound social choice.[184] Hence no meaningful, non-pious domain-restriction can be implicit in contractarianism. De Jasay remarks:

[184] Anthony de Jasay, *Social Contract, Free Ride: A Study of the Public Goods Problem*, Oxford: Oxford University Press, 1990, 100.

> In sum, the conditions under which a social-decision rule would be unanimously preferred or judged indifferent to another require that, one way or another, those called upon to choose the rule should ignore their particular preferences and interests, becoming either superhuman Founding Fathers or subhuman zombie clones of each other. Failing this, the social-choice rule must itself be selected by non-unanimous `social choice'; the logical and moral circularity involved gets no less unsatisfying for being disguised.[185]

Standard-bound contractarianism is inconsistent with the *raison d'etre* of contractarianism: to have social choices made under a rule which selects states of affairs for how well they are *liked* (preference-based rule) instead of for what they *are* (standard-bound rule).[186] I have argued in the preceding chapter for a version of preference-based contractarianism I call *occurrent contractarianism*. In this chapter, I shall explain how contracts are binding for a preference-based version of contractarianism.

There are two issues at stake here. On the one hand, we want to know how it is in an individual's occurrent interest to stick to his agreements when it entails paying debts. The second involves enforcement. How can a model relying solely on occurrent preferences permit enforcement to protect against defectors? I conclude that an enforcement agency is needed to ensure against agreement-

---

[185] de Jasay, 109.

[186] de Jasay, 113.

defectors. Nothing that I say in this chapter makes a commitment on the necessity of a state, however. Enforcement may be supplied by private agencies. And if enforcement is equipped by a state rather than a private agency (or agencies), nothing of what I say here constrains me to any claim about the power of the state beyond the minimal role of contract enforcement. Speculations on these issues are reserved for part four of this thesis.

8.1 PROMISES, DEBTS, AND ASSURANCE

Occurrent contractarianism faces the following difficulty: Contracts remain sacrosanct even if preferences change after the fact. To defend the sanctity of contracts is seemingly to abandon hope of grounding them on individual preference. How then, on the view of occurrent contractarianism, are contracts binding?

Recent attempts at grounding contractarianism on individual preferences have, it seems to me, failed for clandestinely adopting "standard-bound" policies. To do so abandons contractarianism; not saves it. The problem is, can we support a preference-based contractarianism without slipping into standard bound appeals? Claiming that it is in my self-interest, *simpliciter*, to keep a promise seems unsatisfying since in many cases it is not, unless we broaden our "self-interest" to include some notion of what is *really* in my self-interest and admit that this may be

contrary to what actually is in my self-interest. The problem with this has already been stated.[187] We move from preference-based rationales to standard-bound ones. Another possible answer is to tie the moral obligation of adhering to one's compacts by the antecedent moral obligation not to harm another. This too fails. If contracts are binding because of the harm principle, then it would appear that the basis of morality is not contracts, but simply the harm principle. Contracts would be legitimate so long as they are not in violation of the harm principle. If so, contractarianism would be redundant; it would be entirely subsumed under the harm-principle. In point of fact, it is the reverse. The harm principle is not the foundation of morality if what we mean by "foundation" is an underlying *non-moral* reason to be moral. The appeal of not harming another, all by itself, is a moral appeal. How we ground morality, according to contractarians, is by claiming that it is in our self-interest to accept a universalized rule not to harm people. Since we are roughly equal with one another, we would fare worse if harming others were the norm. The chances of our being harmed would be greatly increased, and any net benefit would be greatly reduced. The harm principle, then, is legitimized precisely by what people (would) agree to. It would be circular, then, and of the vicious variety, if we attempt to defend the enforceability of contracts (an aspect crucial to

[187] See my discussion in chapter 4, section 2.

the justification of contractarianism as a whole) based on the harm principle, while at the same time maintaining that the harm principle is itself justified on contractarian grounds.

The harm principle, moreover, cannot limit acceptable contracts because, by itself, it is not sufficient. We need a further criterion to decide what counts as harm. An identical act in identical circumstances may count as "harm" for one group but not another. To forbid certain acts on the grounds that some are believed to be harmed by acts otherwise deemed innocuous would not meet with general agreement. Pornography, for example is considered a "harm" to some groups, but it is non-contractarian to thereby forbid it among all groups. That would be paternalism. To accommodate what counts as harm in a way consistent with how disparate people actually feel about the matter must be decided on an *ad hoc* basis, and this can only be successful if we leave it to the individual preferences of concerned parties.[188] We have moved from the "harm-principle" as being a constraint on contracts, to contracts being a constraint on the "harm principle."

8.1.1 Defaults

[188] I do not advocate that what counts as harm is simply whenever a person *feels* harmed. Should a person feel harmed in a social dealing, she has a right to avoid direct involvement with that dealing, and her non-involvement should be respected. There is a recent trend to define sexual harassment as occurring whenever a person feels harassed. This is clearly inadequate.

De Jasay has claimed, and rightly so, that there are two sorts of defaults and that only one needs to be enforced. In *first-degree defaults*, the unfairness is "benefit-based," that is to say, there is benefit in defaulting. On the presumption that we are all utility maximizers, there is a motive for pursuing any benefit, and hence enforcing adherence to contracts is necessary when benefits from defection is perceived to exist.[189] *Second-degree defaults*, however, do not require enforcement or punishment. This is when all parties to the agreement mutually defect. The end-product is the status quo. No one is worse off, although they are not as well off as they might have been.[190] The distinction is that first-order defects are unilateral defects, and second-order defects are cases where the contract has been mutually annulled.

8.1.2 Spot and Forward Contracts

There is a further distinction important to contractarians which is helpful in seeing the viability of preference-based contractarianism. The distinction concerns the *timing* of the contract. In a *spot contract*, no time elapses between the promise and

---

[189] de Jasay, 34.

[190] de Jasay, 35.

the performance of an agreement, and all performances occur simultaneously.[191] Most ordinary exchanges are spot contracts. For example, giving a clerk money for a pair of boots is a typical example of a spot contract. The performances of each party to the agreement are mutually conditional upon the other party's performance. Spot contracts are intrinsically self-enforcing since neither party to the contract values what he has to give up more than what he stands to get.[192] In a *half-spot, half-forward contract*, one half of the agreement is performed immediately, but the other part of the bargain will be performed at a later time. The second performance is contingent on the first performance, but not vice versa. The second is deferred consideration. The lending of money and the future repayment of the debt is an example of a half-spot, half-forward contract. These hybrid contracts are intrinsically *not* self-enforcing since there is no natural incentive to repay debts. In a *forward contract*, two simultaneously deferred performances are promised. All forward contracts are spot exchanges at a future date. Forward contracts have some tendency to be self-enforcing since we assume the promising is on the grounds of all parties benefiting from the expected exchange. They are not intrinsically self-

---

[191] de Jasay, 22-23.

[192] Otherwise, we assume, the parties would not have entered into the agreement. This does not assume that each benefits *as much* as each other. Nor does it assume there is a unique fair bargaining outcome based, say, on minimax relative concession (cf. David Gauthier, *Morals by Agreement*, Oxford: Oxford University Press, 1986, 143-145.)

enforcing, however, since the future brings with it many unknown intermediating factors. Like the spot exchange, defection from forward contracts will not likely be of the first-order.[193]

8.1.3 Assurance

De Jasay attributes Hobbes's mistake in assuming contractarianism leads to absolute sovereignty on his failing to accommodate well the distinction between spot and forward contracts.[194] If spot-contracts are self-enforcing, then agreements and cooperation are possible in the state of nature; something that Hobbes was unwilling to admit. It is not at all clear that Hobbes was wrong on this, however. De Jasay is right that spot-contracts are intrinsically self-enforcing concerning the possibility of any first-degree default. If you fail to give me the money, I will not give you the boots. But it is not the case that spot-contracts can intrinsically defeat second-degree defaults. We may both be too suspicious of the other to risk performing our part of the exchange. It is true that I may want a pair of boots more than the money in my pocket, but surely I want both the boots and my money more than one alone. But since you can witness my every action, you will not likely be

[193] de Jasay, 23.

[194] de Jasay, 26.

dupe enough to fall victim to any first-degree default in spot-contracts. De Jasay is right about that, but it does not follow that spot contracts are thereby intrinsically self-enforcing. So long as we are roughly equal in character traits, you are likely to prefer having both the boots and the money as well. Or, to put the matter in reverse, above all you would like to avoid ending up with nothing. Me too. Thus, without securing against second-order defaults, it is not clear that we can eradicate ourselves from the state of nature.

To escape the state of nature, it appears we need to enforce against second-degree defaults as well. The main reason for this is to solve an assurance problem, not a prisoner's dilemma.[195] Even if people prefer to make and keep agreements, they will be reticent in doing so if they lack assurance that others will abide by the terms of the agreements. Given this reluctance, and given mutual desires for benefit through cooperative agreements with others, people will have an occurrent interest in removing obstacles to agreements. Enforcement against default will achieve this. So long as there is an enforcement structure against default, it will be in individuals' interests to voluntarily make and keep agreements.

8.2 ENFORCEMENT

[195] For resolutions of prisoner's dilemmas, see ahead to chapters 9 and 10 in this thesis.

That people will voluntarily agree to enforcement is true only so long as the enforcement incurs less cost than the benefit received through agreements. We have seen two ways in which we cannot explain how or why contracts are held sacrosanct. Standard-bound impositions fail the contractarian requirement of being individually motivating. The harm-principle is itself justified by the moral force of contracts, not the reverse. What we need is a positive account of how to make this claim without abandoning our reliance on individual preference.

8.2.1 Formalism

De Jasay distinguishes two approaches: formalism and realism.[196] Under formalism, a contract that is recognized to exist is *eo ipso* binding and entitles the promisee to enforcement. The *adequacy* of the consideration is not a factor. As de Jasay remarks, selling a kingdom for a horse is a valid contract.[197]

Under a realist interpretation, it is not the form of the contract that is binding: it is the *merit* of the case at hand that earn these contracts the socially bestowed

[196] I am not particularly happy about these names, but they will do.

[197] de Jasay, 31.

rank of enforceability.[198] In other words, teleological considerations are what matters in deciding the enforceability of contracts.

De Jasay rejects formalism. It is apparently unfair. "Where neither party has performed and neither has suffered from reliance on the other's expected performance, where is the basis for prescribing a remedy?"[199] De Jasay sees formalism, then, as forcing people to abide by contracts made where all parties have expressly vowed to annul the contract. If this is indeed an implication of formalism, de Jasay is right to reject it. There need not be such an implication, however. Recall the distinction between first and second-degree defaults. Since mutual defection is as having no contract at all (the status quo), enforcing against that is brute state control. Being forced to make and keep an "agreement" is not really agreeing at all. No contractarian, whether a formalist or realist, would accept such an anti-contractarian resolve to the problem of contract enforcement. It is a solution that is certainly not preference-based. Thus, only first-order defaults pose difficulties for contractarians.

If we re-look at what de Jasay finds as intolerable in the formalist position, it is the supposed enforceability against *second-degree defaults*, and not first-

---

[198] de Jasay, 32.

[199] de Jasay, 36.

degree ones, for his fear is enforcing that which "neither party has performed and neither has suffered." If enforceability is attached only to first-degree defaults, then the disadvantage of formalism, at least that identified by de Jasay, dissolves. Formalism, as a principle governing when to enforce contracts, can apply to unilateral breaches only, not mutual breaches. Enforcement against second-degree default need not be a component of formalism. Making such a move, at any rate, is consistent with de Jasay's own distinctions.

The formalist claim is that contracts are enforced when they are "*recognized to exist.*" Typically, contractarianism recognizes any contract so long as all parties voluntarily agreed to it and it entails no relevant externalities. If second-order defaults occur, there is no reason to count that as a contract, rather than as a new contract absolving the first. It is possible, then, to define what counts as a contract formally without making merit-based constraints.

8.2.2 Enforcement

One qualification needs to be made. Enforcement comes in varying degrees from imprisonment to social segregation to mere pitying glances. My discussion of enforcement under the formalist position so far collapses all these sorts together. A further distinction is required to decide between which first-degree defaults merit which degree of punishment. Presumably this depends largely on the costs of entering the agreement itself, as well as on the costs of the varying types of

enforcement. This has a merit-based look to it with one crucial difference. According to a formalist understanding of the justification for contract-enforcement, enforcement is not an external constraint on bargains as much as it is one of the conditions the bargainers place on the contract. Thus, the type of enforcement is as much a part of the contract as any other aspect of the agreement constrained by the preferences of the individual bargainers.

8.2.3 Realism and Utilitarianism

The formalist interpretation is consistent with contractarianism. The realist interpretation, on the other hand, leaves contractarian lines in favour of utilitarian ones. Agreement falls out of the picture in favour of what benefits society as a whole. Following de Jasay, there are three ways of interpreting what counts as "merits," and none accommodate preference-based contractarianism. (i) The first way decides the merits of a contract according to the aggregate social cost-benefit analysis entailed from it. Contracts are enforced so long as they are good for the aggregate society. Not otherwise. Since the ability to rely on agreements is *convenient* for society, it makes good sense for it to help out with enforcement.[200] (ii) The second way comes from recognizing that respect for contracts in general is

[200] de Jasay, 32.

good for society. Hence it is legitimate to induce respect by force. By accepting the benefit of contracting, an individual *effectively incurs a liability* to contribute to its production by helping others keep their agreements, helping victims get recompense, and ensuring the inadmissibility of free-riding.[201] (iii) Alternatively, the realist position may treat each case on its own merits. Contracts would be enforced only if default by the promisor gives him an unfair advantage or constitutes unfair treatment of the promisee.[202]

The third option appears to be the least utilitarian of the three. Merits rest on the harm principle, and this is decided according to the individual situation, not how it relates to society in general. I have already mentioned why grounding the enforceability of contracts on the harm principle is unsatisfactory. The determination of what counts as "unfair treatment" can only be based on some standard-approach, and not on individual preference.[203] If so, preference-based contractarianism must reject it. Alternatively, "unfair treatment" may best be viewed

---

[201] de Jasay, 33.

[202] de Jasay, 34.

[203] Nozick's argument against patterned principles of justice can apply here just as well. Robert Nozick, *Anarchy, State, and Utopia* (New York: Basic Books, 1974) 160-163.

as the occurrence of first rather than second degree default. If it is, this can be accommodated by formalism.

The second merit-based position is an implicit appeal to the problems of free-riding. On the basis that we cannot have free-riders, then everyone must chip in. De Jasay recognizes a problem with this assumption.

> To say that an outcome where free riders benefit from contributions of others is inferior is open to challenge. It could be defended by arguing either that unfairness is sufficiently bad in itself to outweigh the cost of reaching a different outcome, or that in the presence of free riding, contributions will be inadequate. Neither argument is very strong; one of the flaws of contractarianism is to treat them as self-evident propositions.[204]

A more general problem with this second merit-approach is that it too abandons preference as a basis of morality. That I prefer to contribute to x does not in any way necessitate that I also prefer to contribute to y. It raises the problem of "implicit" contractarianism. This is the claim that because I have not rebelled against one part of the state's laws, I thereby automatically accept all of the state's laws. This reasoning evidently was good enough for Socrates, but it abandons the appeal to preferences individuals hold. This version of realism is wholly unrealistic.

---

[204] de Jasay, footnote, 96.

The first merit-based approach appears quite sensible, except for the reliance on the good of the "aggregate society." One of the problems of utilitarianism is precisely its abandoning reliance on individual motivations in favour of the good for society as a whole. This approach needs to explain why I should be more concerned for the aggregate society (to which I for the most part had no choice in belonging) over my own interests. Contrary to Mill, it is not at all obvious why from the premise that "each person's happiness is a good to that person" we can infer that "the general happiness is therefore a good to the aggregate of all persons." If instead, we place the emphasis on individual preferences, this merit-based account for the enforceability of contracts has promise. It would make acceptance of contract enforcement dependent on the individual's interest in having contracts. Making that adjustment, however, removes any distinction it may have had with formalism. It is formalism, and not the realist positions, that recognizes the importance of individual preference and takes as a marker of individual preference any contracts drawn up. To ensure the furthering of individual preferences people would agree to enforce contracts. It is the very nature of contracts, in other words, that they are enforceable. Entering into a contract is not only to accept the particular conditions drawn up, it is also to accept precisely the enforceability of those conditions.

David Schmidtz argues in a similar vein. Contract enforcement can be provided noncoercively. The legitimacy of using force as the instrument of contract enforcement derives from the consent of the involved parties. "An agency that enforces compliance by the principals is doing only what the principals have hired it to do."[205] The enforcement agency collects the debt from the reneger plus an added fee for the enforcement service. The full cost of enforcement, therefore, is on the shoulders of the defectors, not the compliers. No one with sincere intentions to oblige the dictums of their own voluntary agreements need fear the costs of enforcement. If one does not like the attached threat of enforcement, now is the time to withdraw or re-bargain; not afterwards (unless it can be mutual). The formalist aspect of contracts sufficiently answers why contracts are enforceable while at the same time maintaining that their enforceability is not counter to individual preference.

8.3 AGREEING TO ENFORCEMENT

Ideally, what we want to make formally binding are voluntary, part-forward contracts devoid of significant externalities. No external party can step in and say what contracting parties *should have done*. Spot contracts that are defaulted

---

[205] David Schmidtz, *The Limits of Government*, Boulder, Colorado: Westview Press, 1991, 98.

(which we assume will necessarily be second-degree defaults) are permissible, since it leaves all relevant parties in the status quo. It is enforceable that the poor man pays the rich man, so long as the poor man entered the agreement willingly (given his situation). What preference-based contractarians need to explain is how this enforcement is itself preference-based. For the simple reason that it evidently needs to be *enforced* is sufficient indication that it is not in the defector's interest.

The formalist resolve is this: Enforcement against first-degree defaults will make second degree defaults intrinsically unappealing to all parties: and this will be so independently of the timing of the contract. If I enter the contract in the first place, I believe I will do better through the exchange than without the exchange, and the same applies to you. So long as a formal system can enforce compliance, I need not fear your first-degree default, and I have incentive not to commit first-degree default either. Again, the same applies to you. The dominant choice is to comply, and this is so for both of us. Hence, so long as a formalist structure is in place, contracts will be intrinsically self-enforcing. As this will be better for us all, individually defined, it is in our interests to erect a formalized contract-enforcement institution.

8.4 A PARADOX RESOLVED

There is the paradox that we could not agree to erect this formal institution prior to the formal institution itself, and that if we could, then we can make and keep agreements without any such formal structure. This is certainly a problem when what we need is an entire society's agreeing on erecting a sovereign, while at the same time admitting that such harmony among disparate peoples is not possible prior to a sovereign's governance.[206] Fortunately we do not need such wide scale unanimity concerning contract enforcement. Once we accept the formal first-degree default contract-enforcement structure, we are not committed to an absolute sovereign. It does not follow that the sovereign can now command obedience in any domain of our lives. The domain of contract-enforcement is not restricted, true, but the structure does not grant power to the sovereign beyond imposing voluntary, individually chosen contracts to be binding. There is no appeal to a standard in judging these disparate contracts. The content of the contracts, as well as the degree of enforcement for violations of those contracts, are entirely left up to the preferences of the concerned individuals given their situation. The domain of potential contracts is unrestricted. Power remains with the individual, not with a sovereign. The enforcement aspect of agreements is built right in to spot contracts, so we can

[206] This is one of the intractable problems for Hobbes. See Chapter 11 for a discussion of the further problems with Hobbes's political conclusions.

see at least one sort of contract that does not require an antecedent agreement about enforcement. Without requiring this antecedent agreement prior to spot contracts, the paradox does not apply to spot contracts. Under the formalist interpretation of contract enforcement, we can also see that the paradox does not apply to half-forward, half-spot contracts. The first actor has reason to include in his bargain the element of enforcement. Otherwise he risks being made worse off than the status quo. He does not need unanimity about this particular in his proposal among every potential bargainer. He requires agreement only from his co-bargainer. His co-bargainer has no reason to reject the enforcement rider, so long as he has any chance of benefiting from the exchange. Since we assume individuals do not willingly enter contracts unless there is hope of benefit to them, an enforcement rider built into the proposed contract should not involve any cost to those who do not defect. In time, enforcement aspects being implicit in any bargain will become merely part of the convention of contracts. The paradox that is problematic for Hobbes, then, is not problematic for preference-based contractarianism.

8.5 SUMMARY

A number of distinctions among contractarian lines of thought were made in this chapter. To begin with, there are standard-bound and preference-based

versions of contractarianism. This follows from the schism between Locke and Hobbes. Locke introduced standard bound restrictions on the sovereign's power. Although we want to restrict Hobbes's sovereign, I have argued that this can be done while remaining true to preference-based contractarianism. The problem with preference-based versions, however, is that they are difficult to support the needed enforcement of contracts. To put this in Hobbes's terms, reliance on mere preference-based contractarianism without an absolute sovereign makes it difficult to escape from the state of nature. De Jasay believes this problem can be circumnavigated by appealing to a realist recognition rule. It is the *merit* of the case at hand that earn these contracts the socially bestowed rank of enforceability. Realist versions for how contracts are binding are utilitarian, however. As such, they appeal not to individual preferences, but some conglomeration of preferences. As these cannot be assessed by appeal to individuals, for that would be measuring individual preferences and not conglomerate preferences, the appeal is to standards of social import. Thus, realist interpretations abandon the underlying core thesis of contractarianism: the appeal to individual preferences. I have defended, instead, a formalist account of the enforceability of contracts, supporting this by the degree to which it furthers individual preference.

"Occurrent contractarianism," as I have called it, is rooted in our social-psychological state. Given the characteristics we have, and given the social situation in which we are embedded, the best resolve we have of furthering our individually defined preferences is to adopt and adhere to a moral system. Occurrent contractarianism remains true to the original contractarian insight; that morality is a rational institution, capable of being designed for and adhered to even by non-tuistic rational beings following merely their own occurrent preferences.

Chapter 9

SELF-EFFACEMENT

I have remained adamant that self-interest is the sole guiding force of our moral motivation. This means, crudely, that it is in our self-interest to be moral. This certainly follows if we understand self-interest in non-hedonistic terms. We can see clearly that the addict's life would go better by undergoing a painful period of self-effacement. After the transformation from addict to non-addict, he will himself, undoubtedly, appreciate this as well. Let us grant all that. The trouble is, I have adopted a hedonistic account of self-interest. The non-hedonistic account, I have argued, is a standard-based view of self-interest, which I reject. How then, are we to deal with the addict. How are we to deal with the adamant rapist?

Dismissing the notion of "considered preferences" leaves contractarians with the problem Gauthier, Frankfurt, Butler, Baier, Taylor, and Schmidtz were individually trying to solve. Surely morality cannot be grounded on self-interest if self-interest includes acting on whims and impulses that will counteract self-interest in the long run. Acting on short-sighted self-interest is precisely what makes the prisoner's dilemmas so intractable. The moral choice, we know, is

cooperative behaviour. If there are no restrictions on individual preferences, dim-sighted individuals may defect from prisoner's dilemmas, and it would be silly to claim therefore defection is the moral choice. We can see that cooperation is better for the individuals than mutual defection. But, according to my analysis above, claiming that therefore cooperation is what is "really" in the players' interests is to introduce standard-bound or objective criteria, and doing so abandons individual preference entirely.

Let us imagine that Joe (not his real name) has an occurrent preference to rape a particular girl. Given this preference, a particular course of action may be planned that would ensure his success. Approach her stealthily from behind, hit her over the head, threaten her further with a knife, etc. The satisfaction of Joe's occurrent preference through this stipulated action indicates its rationality. Since morality, as we've so far described it, is at least partly a rational act satisfying an occurrent preference, it would appear that the rape of the girl is morally permissible. It is not, of course. It is immoral not merely because we know rape is immoral in the real world, for at the level of theory inception, we have as much reason to change our current moral intuitions as the theory, should conflict between the two arise. At least we make no presumptions in favour of folk-ethics.[207] In this case, however, there is no conflict; theory and folk-

[207] This goes against Rawls's method of reflective equilibrium (see John Rawls, *A Theory of Justice*, Cambridge, Massachusetts: Harvard

ethics coincide. The rape can not count as moral on occurrent contractarian grounds. Recall that for contractarians, what is moral is not simply what satisfies an individual's occurrent preferences, but what also does not interfere with relevant others' occurrent interests. We shall assume that rape is not in the interest of the girl. Since she is a relevant other in this scenario, her objection has weight enough to stymie our calling Joe's act moral.

"True enough," Joe may say, "but who cares. I agree it is not a moral act according to your theory, but why should that daunt me? Look, you've told me about the importance of pursuing one's occurrent preferences; well, here is my occurrent preference, so I'm going to pursue it. Your own theory tells me that no one can get on a high horse to tell me what occurrent preferences I ought to have." If my theory has nothing to say to this man, it fails its task. So in this chapter, I explain how occurrent preferences are key, but why nevertheless, self-interested rational individuals should wish to alter some of their occurrent preferences when carrying them out violates the contractarian credo. To help me in this task, I will appeal to recent works in game-theory.

9.1 A GAME THEORETIC APPROACH

University Press, 1971, 48-51.

Recent work in game theory, starting from David Gauthier's seminal work in *Morals by Agreement*,[208] and developing further in Peter Danielson's work, *Artificial Morality*,[209] has attempted to show how moral dispositions win-out over immoral dispositions through repeated prisoner dilemma situations. From this it is speculated that morality, being the rational choice for survival, is a successful evolutionary strategy for homo-sapiens.

There is much in these works, I believe, but there is also, it appears to me, much fudging. The issues are complicated, the limiting assumptions contentious, the variables many, and the values plugged in to the formulas appear arbitrary and *post hoc*. In this chapter, I wish to raise the difficulties with game theoretic approaches, without yet disparaging the whole project. Failing hard and fast solutions, my conclusions here are tentative. I gain some hope for their being coincident, at least, with Gauthier's and Danielson's central conclusions: that it is rational to change one's stripes.

9.1.1 Gauthier's Model

---

[208] David Gauthier, *Morals by Agreement*, Oxford: Clarendon Press, 1986.

[209] Peter Danielson, *Artificial Morality: Virtuous Robots for Virtual Games*, London: Routledge, 1992.

Gauthier and Danielson agree that the people (or entities) with cooperative dispositions do better than those without. Doing better typically means getting higher scores in an interactive prisoner's dilemma (PD) game. The moves of the game are to cooperate or to cheat. It is thought that successful cheating on your part yields the most points for you. Keeping at merely an ordinal ranking, we will give the successful cheater 4 points. Cooperation is good too, but not as good. Typically cooperation is better than the status quo, otherwise we would not waste our time. Cooperation, then, gets 3 points, while the status quo, which in these games is unsuccessful cheating or mutual defection, gets 2 points. The worst outcome for a player is to be cheated. That means, while you play the cooperative moves, your opponent cheats on you. Being cheated is worse than the status quo, and thus gets only 1 point.

In a two by two matrix, the possible outcomes look like this:

		A play fair	A cheat
B play fair		3 3	4 1
B cheat		1 4	2 2

Table 9.1. The Prisoner's Dilemma (PD) Matrix

A's outcome values are shown in the upper right corner of the cells, and the outcome values for B are in the lower left hand corner. The problem, one can see, is that if B were cognizant of his expected outcome values in this game, he would cheat. That is, no matter what A does, plays fair or cheats, B will do one better by cheating than by playing fair. Let us imagine that A plays fair. If B plays fair, he gets 3 points, which is good. But if he cheats, he gets 4 points, which is better. So, if A plays fair, B will cheat. If, on the other hand, A cheats, then of B's two responses (play fair or cheat), again B's best response is to cheat. Cheating, when A cheats, gives B 2 points rather than 1 point had he played fair. The result is that it is rational for B to cheat, no matter what A does.

This is an awkward conclusion for those of us who wish to show that it is rational to be moral. What the prisoner's dilemma seems to show is the reverse: immorality is what is rational. But wait! If B can reason in this way, so too can A. Thus, in the same way that it is rational for B to cheat, the dominant strategy for A is to cheat as well. What this means, in this game, is that A and B will each get 2 points. In other words, they will each and seemingly forever remain at the status quo. When there is a strategy that could leave them both better off (both playing fair, and each getting 3 points), it makes one wonder if A's and B's reasoning really

is all that rational. Recall, rationality is a tool to yield the most (or an adequate number of[210]) utiles possible, where utiles are scores of preference satisfaction.

David Gauthier reasoned that if we take two types of players and run a number of the PD games with them, the type of player that is more willing to play fair is going to win out (in other words gain more points) over one who is going to cheat whenever he can. For this outcome, a limiting assumption needs to be made. As put so far, anyone who plays fair against a cheater is going to do far worse than that cheater. The cheater will get 4 points in that interaction, while the cooperator (as we shall now call the person who plays fair) will get only 1 point. Although aware of this, Gauthier believed that so long as the intentions of the players were known before-hand by the other players, then it would be wise to cheat with those you knew were going to cheat and cooperate with those you knew were going to cooperate. There is a sense in which we bring in Newcomb's paradox to get this to work. At face value, it would seem that if you knew your opponent was going to

---

[210] The debate between whether rationality should be viewed as preference *maximization* or simply preference *satisfaction* is still raging. To claim that an act is rational only if it *maximizes* preferences may seem too strict. On certain occasions I may be perfectly content to merely *satisfy* my preferences without absorbing the cost of checking to see whether another choice could improve my preference satisfaction. Seeking to always maximize preferences has the drawback of calculation and time costs that a preference satisfaction model avoids. A problem with the preference satisfaction model, on the other hand, is that it seems to overly glorify the status quo. See David Schmidtz's discussion in "Rationality within Reason," *The Journal of Philosophy*, 89, 9, 1992, 445-466.

cooperate, it would be better for you to cheat (you would get 4 points rather than 3 points). The assumption prevents this, however, for your opponent's intention to cooperate is dependent on your intention to cooperate. If you now change that intention to an intention to cheat, your opponent now knows this, and so too, he will cheat. Thus, the only way you can gain the maximum points is for you to adopt, first off, an intention to cooperate with other cooperators. Your opponent now sees this, and if he is like you, which is also assumed from the assumption of equal rationality, he will form the intention to cooperate as well. *Voila*, you each get 3 points. This, note does better than the person who decides to cheat regardless of the other's intention so long as other players have only the intention to cooperate with other cooperators and not the intention to cooperate unconditionally. Given the condition of the transparency or omniscience of players, cheaters can expect only mutual defection, not unilateral defection, and thus will receive only 2 points.

Gauthier has names for these different players. A Straightforward Maximizer (SM) will cheat in all cases. He reasons that whatever the other player does, he does better to cheat. A Constrained Maximizer (CM) forms the disposition to cooperate only with other cooperators. CMs will not fall prey to SMs, for CMs will cheat with SMs. And for interactions with other CMs, they each will reap the rewards of 3 points rather than the 2 points that SMs are destined to earn.

The conclusion is that over repeated games (with many players) CMs earn more points than SMs. Thus, it is more rational to adopt the CM disposition than the SM disposition. The moral of the story is that this parallels real life; people who cooperate with others who cooperate do better in the game of preference satisfaction (or maximization) than those who do not adopt this disposition. Morality is the rational choice.

Some have disparaged Gauthier's tale. What is this notion of being able to accurately predict the other's strategy? In real life we are not dealing with the aliens of Newcomb's problem. I can only know what my opponent's *voiced* intentions are. How do I know my opponent is telling the truth? If rationality is not to maximize but to disaster-avoid, then perhaps it is rational to cheat even if I have the belief my opponent is telling the truth about his intention to cooperate. Moreover, not all interactions are what de Jasay calls *spot contracts*.[211] More often than not game situations in the real world are part-forward part-spot contracts. If he's already given me the money for services rendered tomorrow, why shouldn't rationality tell me to cheat in this case? I am assured the 4 points. Why settle for only 3? In brief, Gauthier answered these sorts of objections by claiming that in the real world, *reputation* matters. True, I may get 4 points now in this encounter, but I will be in

[211] Anthony de Jasay, *Social contract, Free Ride*, Oxford: Clarendon Press, 1989, 22-23.

effect branded with a mark that tells my future opponents that I am an SM, and not to cooperate with me. If I fear this, as I should, since it has been shown that SM's do worse than CMs, then I have reason to constrain my motivation to maximize on particular occasions for the reward over time. Based on the very real aspect of reputation, as well as the psychologically revealing body language of our species, this is all we need to indicate what our intentions are to others. We need not be transparent, Gauthier claims, but we are, in fact, *translucent* enough to favour the CM disposition over the SM disposition.[212]

A more sophisticated complaint about Gauthier's model is that it is recursive, and hence unworkable. I will cooperate with you only on the condition that you will cooperate with me. But meanwhile, you will cooperate with me only on the grounds that I will cooperate with you. We seem to be at a standoff. It is more likely that we will both defect than that we will both cooperate. On a computer program, for example, the computers would not know what to do. A's program says: cooperate with B only if B cooperates with A. But B is not cooperating with A right now. Instead, it is waiting for A to cooperate with B. But A is not doing that, because it is waiting to see if B is going to cooperate with A, etc. etc.

---

[212] Gauthier, 174-177.

9.1.2 Danielson's Pluralistic Model

Peter Danielson agrees with the main thrust of Gauthier's work, but argues that it is not complete.[213] There are more dispositions than just two. For example, to understand Gauthier's CM, we needed to contrast it with another disposition, a disposition that cooperates with whoever whenever. Danielson calls this disposition an Unconditional Cooperator (UC). Gauthier's CM cooperates with both CMs and UCs, since Gauthier's CM has the program to cooperate with other cooperators. Because CM is a broad category for dispositions that constrain their maximization programs for the better, Danielson renames Gauthier's CM as Conditional Cooperator (CC). This CC cooperates with other cooperators. Thus, CC will cooperate with other CCs as well as with UCs. Meanwhile, UCs will cooperate with other UCs, CCs, and even, foolishly, with SMs. Because UCs cooperate with SMs, even though they know they will be cheated as a result, this will in turn benefit SMs, who will start doing better. CCs get 3 points for every interaction with CCs and UCs, while SMs get 4 points for their interactions with UCs and 2 points for their interactions with CCs and other SMs. Although SMs do not do as well as CCs with the introduction of UCs, they do better with UCs in the population than without for obvious reasons.

---

[213] Danielson, 12-16.

This is not Danielson's stopping point. With the introduction of UCs, a more rational disposition emerges: what Danielson calls Reciprocal Cooperators (RC). Unlike CCs, who cooperate with other cooperators *simpliciter*, RCs cooperate only with those whose cooperation is dependent on their cooperating. RCs differ from CCs in that, although CCs will cooperate with UCs, RCs will not. They will not because UC's cooperation is *independent* of RC's cooperation. Since UC is going to cooperate anyway, why settle for only 3 points out of an interaction with them, when one can, without cost, get 4 points. Cooperating with UCs simply makes no sense. Danielson's RC disposition does better than either the dispositions of CC, SM, or UC. For where SMs and CCs get an average value of 3 each for every encounter with either UCs or CMs (2 and 4 for SM and 3 and 3 for CCs), RCs do one better. They get 3 for their encounters with CCs and 4 for their encounters with UCs. The superiority of the RC disposition is not altered by including interactions with RC. Both RCs and CCs get 3 points for their interactions with other RCs.

A formal account of the scores looks like this:

RC = 4(u)+3(c+r)+2(s)-3
CC = 3(c+r+u)+2(s)-3
SM = 4(u)+2(c+r+s)-2
UC = 3(c+u)+(1)(s+r)-3

The rationale here goes as follows: The lower case letters represent the population size of each disposition in the game. "c" is the population of CCs. "r" is the

population of RCs. "u" is the population of UCs, and "s" is the population of SMs. If a CC meets another CC, we expect 3 points. If the CC meets an RC, again 3 points. If that CC meets an SM, then only 2 points. That CC would get another 3 points for an encounter with a UC. But since he cannot compete with himself, we subtract the 3 points he would have made. Now if there are 10 CCs, 5 RCs, 3 SMs, and 2 UCs in this population, the score for each disposition would be:

RC = 4(2)+3(10+5)+2(3)-3 = 56
CC = 3(10+5+2)+2(3)-3 = 54
SM = 4(2)+2(10+5+3)-2 = 42
UC = 3(10+2)+(1)(3+5)-3 = 41

The moral in this case is not as clear cut as Gauthier's. In Gauthier's there was a simple parallel to be drawn between doing better in the two-disposition PD and being moral in the real world. Being moral is to cooperate and to retaliate against non-cooperators. But Danielson has shown that it is not precisely the nicest guys that finish first. The rational disposition is RC, but this does not quite map onto what we normally consider moral in the real world. We do not think that it is moral to rob Mother Theresa simply because she's a do-gooder.

9.1.3 Chicken

In the game of chicken, two contestants race their cars toward each other. The winner is he who swerves away last. I win if I keep straight while he swerves. Next

best for me is if we both swerve at the same time. That would be a tie. Third best is if I swerve and he keeps straight. Although I remain unharmed, I lose the game. The worst outcome for me, however, is if we both keep straight and our cars ram into each other and we die or suffer terribly. A decision matrix for the chicken game, thus, looks like the following:

		A: swerve	A: straight
		2	1
B	swerve	2	3
B	straight	3	4
		1	4

Table 9.2. The Chicken Game Matrix. (4 least favoured; 1 most favoured.)

The dominant strategy is for us both to swerve.[214] But then the dominant strategy of the game is not to win. This is peculiar. Of course, swerving *at the same time* is difficult to do, even if it were coordinated. Given this, one could hope to swerve slightly after one's opponent. This would be, in effect, keeping straight and winning. Of course, assuming both contestants are equally rational, the other contestant will be thinking the same thing. The result is that both will be waiting for

[214] Danielson argues there is no dominant strategy, but that is because he rules out simultaneity (Danielson, 166).

the other to swerve first, and thus, both will swerve too late. It appears that self-interested, instrumental rationality dictates the worst outcome, and not the dominant strategy in the game of chicken.

In real life, however, we have a decision to make prior to this matrix. The dilemma comes only once we have already committed ourselves to play. And this presupposes a certain character trait that I think most rational adults do no not have. The decision matrix to make prior to engaging in the game of chicken is represented in table 9.3 (on the following page). Given that playing the game of chicken is likely to yield the worst outcome for equally rational players, it can be determined in advance that playing the game is worse than not playing the game. True, A would save face if B also did not play. A does not "lose," though, because he did not enter the matrix of the game which is required for there to be a loser. This slight embarrassment, at any rate, is petty compared to the expected outcome from playing. This is one problem with these matrices, that they cannot adequately represent cardinal ranking. A's declining to play does not, all by itself, preclude B from playing. B can always try to find some other sucker. Thus, given the slight disutility in declining when the other wants to play, it is not quite the case that whatever B does, A would do better by not playing. B, knowing that A will not want to play, may broadcast his request to play, even if B himself has no intentions on playing. Notwithstanding the annoyance of dealing with someone such as B, given

the great negative weight of playing and crashing, it is better for A not to play. Since we assume equal rationality in A and B, B should be reasoning the same way. The dominant strategy is not to play. And this shows that the problem with the game of chicken is that it already assumes irrational players.

		A: play	A: don't play
B	play	4 4	3 2
B	don't play	2 3	1 1

Table 9.3. Payoffs for playing the Chicken Game (4 = least favoured; 1 = most favoured.)

This analysis does not change if the object of the game of chicken is not to win, *per se*, but to impress sideline spectators. In such a case, it would seem that the rational move in the game of chicken is to have both players coordinate prior to the game when each player is to swerve. But so long as B is going to stick to the prearranged point of swerving, it would be better for A to swerve after that point in order to impress the spectators the most. Since we assume B is as rational as A, B will be thinking much the same thing, and A should realize this. Should there be any suspicion that the other player will swerve later than the pre-appointed point, it will be more rational not to play rather than to play. Whether

or not the spectators are actually impressed by chicken games, to try to impress someone by the game of chicken is a sign of irrationality, given any desire to survive. Declining is the rational resolve.

The literal game of chicken, then, is not a problem for rationality.

9.1.4 The Threat Game

This analysis goes too fast, however. The chicken game models threat situations. It is precisely the point that since it is irrational for you to engage in chicken games that it becomes rational for me to threaten chicken games for my own profit. Because you see it is in your best interest not to play, I can use this to my advantage. Let us imagine, rather than the chicken game, that I have strapped to my belt explosives, and I threaten to set these off in your vicinity if you do not give me $1,000. You cannot merely refuse to play in this situation. That is to be interpreted as calling my bluff. The result for failure to play, then, may be your death. Deciding whether or not to play is not an option for the threats the chicken games are attempting to model. Threat exchanges are situations in which it costs you to fail to play. You are held at ransom. But since it costs you more to play (risk of death), I can count on your paying me the ransom for not playing. Your rationality dictates it. The problem, we see, is that

rationality encourages bullies in these threat games. It pays to threaten. This is a problem for rational-based morality. What is the solution?

Acting on threats are understood in chicken games to be not only disadvantageous to the threatened player, but also to the threatener. It is not a threat, otherwise. It is not a chicken game. You must intend to do an action it does not advantage you to do. This fact makes it possible to respond to our question above. Acting on one's threat is not rational. If you call my bluff, I have a choice: to follow through on my threat or not to. According to the matrix below (table 9.4) following through on my threat intention gives me my worst outcome. Chickening out gives me a better outcome. Hence, rationally, I should chicken out.

		A: act	A: chicken
	B: act	4 4	3 1
B chicken		1 3	2 2

Table 9.4. The Threat Matrix. (4 least favoured; 1 most favoured.)

It follows that if you threaten me, it is rational for me to call your bluff. It will be rational for you at that point to chicken out and I get my best outcome. If you can predict this outcome, then it was not rational for you to threaten in the first place. If we cannot rationally carry out our threats, it is irrational to threaten. We see that

entering the game of threat, no less than entering the game of chicken, is irrational at the outset.[215]

I have moved too fast once more, unfortunately. Insincere threats against gullible players may be quite prosperous. I may have no intention on carrying out my threat, but so long as I can convince you that I am serious, you will not likely risk calling my bluff. Moreover, if I can actually alter my disposition to in fact enact on called threats, and publicize this, then my victims need no longer be merely gullible players. If self-effacement achieves my desires, rationality will in fact dictate this character transformation.[216]

To successfully threaten someone, you must make them believe that you will, in fact, carry out your threat. Since it is patently irrational for you to do so, you must convince the other player that you are, in fact, irrational. If I know I am encountering an irrational individual, obviously I can not trust that he will do the rational thing. If you threaten me and convince me that you will carry out your threat even if it is against your interest, then I may do better by not calling your bluff. If you can

---

[215] Gauthier argues along these lines. "If it is not rational to act on one's threat because of the costs to one's life going as well as possible, it must not be rational to sincerely make the threat in the first place" (David Gauthier, "Assure and Threaten," *Ethics*, 104, 1994, 719). But Gauthier fails to consider the problems with this response that I raise below.

[216] See, for example, Duncan MacIntosh, "Persons and the Satisfaction of Preferences: Problems in the Rational Kinematics of Values," *The Journal of Philosophy*, 40, 4, 1993, 177.

convince me, in other words, that the top right hand corner of the threat matrix (table 9.4) is an impossibility, then if I call your threat I am committing myself to accept my worst outcome, given your irrationality. But since I can do better by paying you off for my not playing, it is irrational for me to call the bluff of a madman. Madmen, then, do very well.

Can you convince me that you are irrational? Quite likely not everyone will be able to. They will have to suffer along with the rest of us without resorting to threats to get our way. The problem of threat remains, however, so long as self-effacement to a disposition conducive to irrational sincere threats is conceptually possible. Is it conceptually possible? The claim is that sometimes it is rational to be irrational. If whatever furthers your preferences is rational, and being x (in this case being irrational) does this, then being x (in this case being irrational) is rational. This is peculiar, since under this guise, being irrational is being rational, and hence, it is not the case that one ever gives up one's rationality.

What if I reason in this way, and therefore call your bluff? What are you to do? So long as you did not somehow hard-wire yourself into acting on your threat, it will be rational for you to back down. But is it possible to hard-wire yourself in the requisite fashion? And if it is possible, is it rational to do so knowing that you will be unable to back down when it is in your interest to do so?

Let us say a pill is invented that, when swallowed, produces this hard-wire effect.[217] So long as you broadcast that you have taken the pill, this should deter anyone from calling your bluff. Thus, you should have no need to ever back down yourself. After considering this, you take the pill, and will irrationally stick to your guns. You will blow us both up if I don't give you the $1,000. Assuming this is possible, might we also assume I can take a similar pill? It would not be to my advantage if I wait to transform myself only after you do so first. Two unconditional threateners competing against each other in the threat game will do very badly indeed. Perhaps, however, precisely to avoid being taken advantage of in threat games, I have taken a threat-enforcer pill prior to your arrival. Now I am hard-wired to always call someone's threat; and I publicize this fact. If you knew this, you would not target me. I can thereby successfully avoid threats.

Barring the annoyance that we are leaving reality behind in this answer, there is a further problem; the invention of yet another pill. Nothing prevents a further disposition being somehow fashioned that ignores all unconditional bluff-callers. The taking of this pill is also, of course, broadly advertised. If enough of these absolute threat masters exist in the population, it will no longer be rational to adopt the unconditional bluff-caller disposition.

---

[217] I owe this discussion to Paul Viminitz.

In the above speculations, we have immersed ourselves too much in the artificial world. In the real world, there are psychological, biological, and sociological constraints on the kinds of dispositions we can successfully adopt. This means, unfortunately, that we may acquiesce to insincere threats. And doing this makes it rational for more insincere threateners to evolve. What is not accurately portrayed in the ordinal ranking of the matrices above, however, is that in threat games, mutual defection (where threats are acted upon) are exceedingly costly. In many cases, the cost is far greater than the benefit from success. This needs to be factored into the formula before one can decide whether to become a sincere threatener. There will be some who decide the risk is worth it. To alter our own dispositions to counteract such defectors will quite likely be too high a price to demand. That means, unfortunately, that we must be willing to tolerate some degree of threat; and that this toleration itself creates a niche for others to capitalize on the perceived benefits of threat. It is difficult to say, at this point, where the line of toleration ends, and what precisely it is that keeps the numbers of sincere threateners low. Is it that the threshold of retaliation is itself low, or that the biological and social-psychological factors are at work to make the necessary adjustments toward irrationality implausible? I agree with Danielson that rational

moral players must be willing to tolerate some unfairness in the world simply to avoid costly sanctions.[218]

9.2 RESULTS

This does not quite mean we have no definitive answer to give to our troublemaker, Joe, introduced at the beginning of this chapter. Even if it is the case that Joe is in such a condition that it is rational for him to adopt a cooperative disposition over a defecting disposition does not mean that Joe will of his own choose not to rape that poor girl. We can say it is rational for him to refrain from rape even by his own estimations, as spelled out through game theory analysis and given certain specified variables, but he may nevertheless decline to listen. In such a case, why do we have a right to intercede? It would go against occurrent contractarianism if this right to intercede in Joe's libidinous pursuit is grounded in what is in Joe's best interest according to *our* conception, and not his own occurrent interests, however deranged. The right to intercede then, comes from where?

The answer is that to violate Joe's occurrent interests in this case is in *our* interest *simpliciter*. Recall at least some of what we learned from game theory: it is wise to cooperate with other cooperators, but not wise to cooperate unconditionally.

---

[218] Danielson, 194.

We are no dummies; we refuse to cooperate with other cheaters. Joe is a cheater, and therefore it is wise not to do business with him. Nor, if we can help it, do we allow Joe to do business with us. Whatever his occurrent preferences are, he is not one of us. Just as we should not flinch if a CM does not cooperate with an SM, we should have no pangs in violating Joe's occurrent preferences.

The formalist constraint on contracts and agreements help not only to assure fair dealings with one another, but it also acts to rid us of the defectors.

Our game analysis, although not yet complete, has taught us that it is not rational to be a UC and we will not forget that lesson now in our dealing with the Joe's of this world. For this reason, it is quite consistent for occurrent contractarians who preach general tolerance of occurrent preferences to nevertheless intercede in cases where the furthering of an individual's preferences violates the preferences of another. Recall, morality enters the picture in cases of interpersonal relations. Maximizing occurrent preferences will necessarily have to be tailored to fit with the occurrent preferences of one's partner. Rationality says: maximize your occurrent preferences. Morality says: Yes, so long as they concur with related others's occurrent preferences. And game theory adds: true, but not to the point where we no longer satisfy our occurrent preferences at all.

The rapist violates the victim's occurrent preferences. That society should care and intercede is in the interests of the citizens should they desire to live in

peace. That we can force defectors to curb particular occurrent preferences under occurrent contractarian theory is because we are not so irrational as to allow defectors to harm us. A morality consistent with occurrent preferences does not demand that we be a society of UCs.

There is one other important lesson game theory has taught us. It is not so much that we rationally ought to change our occurrent preferences from immoral or amoral to moral ones, but that evolutionary processes naturally favour selection of moral occurrent preferences. Morality is evolutionally successful.[219]

[219] Danielson has been exploring evolutionary models of moral behaviour recently. See Peter Danielson, "Evolutionary Models of Cooperative Mechanism: Artificial Morality and Genetic Programming." To appear in P. Danielson (ed.) *Modelling Rationality, Morality, and Evolution*, New York: Oxford University Press, 1995. See also Peter Danielson, "Evolving Artificial Moralities: Genetic Strategies, Spontaneous Orders, and Moral Catastrophe," read at the conference on "Chaos and society" at the Université du Québec à Hull, June 1-2, 1994.

Chapter 10

COMPLICATIONS FOR GAME THEORY

The account of game theory in the preceding chapter is still too elementary. There are a number of things that it does not take into account. It does not take probabilities of accurate appraisal of one's co-player into account, for example. Instead, Danielson's depiction assumes for simplicity transparency among competitors. It also assumes ordinal ranking only. The outcomes will be different if we can somehow take into account cardinal ranking of the outcomes for the players. It is true that I may value successful cheating over cooperating, but by how much? And will this neatly map onto the cardinal values my competitor places on cheating and cooperating? We know, for example that prison is a deterrent for middle-class people, but it does not appear to carry the same value for the destitute living on the streets, or for members of gangs in which prison records are a status symbol. Another problem is the lack of a significance test. Should I clearly value 56 points over 54 in the above scenario? Under what conditions should differences of such slight proportions matter? Moreover, we will get different outcomes depending on different distributions of the other dispositions. A lone CC in an SM population does

no better than the SMs. An SM in an otherwise pure UC population does extremely well. How are we to calculate the population distributions of the varied dispositions to successfully link game theoretic results to real world applications? Lastly, some cost must be applied to the extra scrutiny necessary for RCs and CCs compared to the unconditional responses of SM and UC. What criteria can we apply that will adequately translate to our real world situation?

Let us look at these complications one at a time.

10.1 TRANSPARENCY

Danielson assumes "transparency" for his model, as well he might, since he is working with virtual robots rather than people.[220] If I am an RC, I can make the distinction between encountering a UC and a CC.[221] Failing this, RC can not be shown to be a rational improvement over CC. As people are *not* transparent, however, it is not wise to assume transparency in order to get us out of self-defeating PD situations if the goal is to show why people developed morality (or would develop morality) as a rational response to PD situations. True, this was not

---

[220] Peter Danielson, *Artificial Morality: Virtuous Robots for Virtual Games*, London: Routledge, 1992.

[221] RC = Reciprocal Cooperator. UC = Unconditional Cooperator. CC = Conditional Cooperator. For an account of these dispositions, see my chapter 9, section 1.2.

Danielson's goal, but it is mine. We need, then, to alter the payoff formula slightly by introducing Gauthier's more realistic assumption of "translucency."

To account for translucency, we will require probabilities in the payoff formula. The difference between transparency and translucency is that one's reading of another's disposition under transparency conditions is 100% accurate. It is less than 100% accurate for translucency conditions. It must be more than 50% accurate, however, since that is no better than chance. Setting the precise probability of success is somewhat arbitrary, I'm afraid. Moreover, some people may be better at it than others. An interesting exercise, in fact, is to wonder how much one's detection abilities determine what disposition one should rationally assume. For example, although an RC may clearly do better than a CC, I may be unable to successfully distinguish between the variety of cooperating dispositions. UCs, CCs, and RCs may all look alike to me. If I do no better than chance in distinguishing between CCs, UCs, and RCs, the most rational disposition I can adopt is CC. Laying aside this worry of how to properly account for the level of translucency, for now let us say the translucency condition is 80% accurate.

There is some reason to suspect we do better than this for we are not fooled in 20% of our encounters. A number of things can be said in defence of setting the translucency level at the 80% mark. That we may not be fooled in 20% of our encounters is not evidence that we are not mistaken in 20% of the encounters. For

one, as will be evident presently, some mistakes do not matter. Also, wary of this 20% error rate, we may decide not to engage in some games with some people, preferring to play with those we have more assurance will play fairly. If so, our success rate of detection is based on a biased sample. Lastly, there are a number of studies that show how easily people are influenced by irrelevant information. For example, attractive people are given the benefit of the doubt.[222] Favourable actions of attractive individuals are attributed to the individual themselves. Unfavourable actions of the attractive people are blamed on external factors such as the situation, other people, or some accident. The reverse is the case with unattractive people. Bad acts are thought par for the course for them, and good acts are a result of external factors. On the assumption that looks have little to do with disposition, mistaken identity would likely be greater than 20%.[223]

Although I can not unequivocally state that an 80% chance of mistaken identity is an accurate appraisal, we need to agree on some weighting of which any

---

[222] Karen Dion, "Physical Attractiveness and Evaluations of Children's Transgression," *Journal of Personality and Social Psychology*, 24, 1972, 207-213. Thomas Baglan, "Effects of Interpersonal Attraction and Type of Behaviour on Attributions," *Psychological Reports*, 48, 1981, 299-304.

[223] A complication is the reasonable supposition that we are who other people think we are. Being treated like a crook may well make us a crook, and likewise, being treated as a good guy may well make us behave that way as well. See for example, Mark Snyder, Elizabeth Decker Tanke, and Ellen Berscheid, "Social Perception and Interpersonal Behaviour: On the Self-Fulfilling Nature of Social Stereotypes," *Journal of Personality and Social Psychology*, 35, 1977, 656-666.

encounter in our game-theory tournaments will have to take account. One might expect this is easily done by multiplying the population of any one disposition by whatever the probability of successful detection we assign (for now on, we assume .8). Although this approach lowers the total scores, it does not alter the ranking. The formula for CCs scores would become 3[c(.8)+r(.8)+u(.8)] + 2[s(.8)] - 3,[224] which equals 38.6, still slightly lower than RCs altered scores at 39.4. But it is not that easy. For if an encounter with a CC normally would give another CC 3 points, we need to take into account the chance of its being specifically with an SM instead. Only in the chance that the CC unilaterally cooperates with an SM, will the CCs scoring for that encounter differ. Under conditions of translucency, for every CC encounter with whom he believes to be another CC, there is a 20% chance of this encounter being with a non-CC. Since a non-CC could be either an RC, UC, or SM in our scenario, we assume a 6% chance of its being with an RC, a 6% chance of its being with a UC, and a 6% chance of its being with an SM.[225] Once we recognize this, notice how this now favours SMs 6% more for every encounter with a non-SM. That is to say, although CCs and RCs stand to be dealing with an SM without their

[224] For an explanation of this formula, minus the .8 translucency condition, see chapter 9, section 1.2.

[225] Assuming equal distributions of RC, UC, and SMs in the population. Complications arising from violations of this assumption are discussed in the following section.

knowledge 6% of the time, while they lose 6% of these encounters (getting 1 rather than 3, or for the RC's belief of dealing with a UC, 2 rather than 4), the undetected SM stands to gain 4 rather than 2. Thus, by introducing the translucency condition, SMs stand to benefit. And the lower the percentage of accurate detection of dispositions, the better the odds favour the SM disposition. For example, under conditions of complete opacity, the risk may be too great to cooperate in any encounter,[226] thus making the SM disposition attractive. Under translucency conditions, the SM disposition will be more rational than under conditions of transparency. Importantly, the RC disposition fares worse under conditions of opacity.

Danielson was not unaware of this, mind you. He did recognize that under conditions of translucency, rather than transparency, RC will likely pay higher costs than CC.[227] He did not attempt to provide a formula for this, however. Instead, he assumed the following table of outcomes would be likely after the appropriate scrutiny costs have been incurred:

---

[226] I say the risk *may* be too great, rather than is too great, because it matters what constitutes the overall population. If somehow we know everyone is a CC or UC, there is little risk in cooperating. Although in a population of CCs and UCs and under conditions of opacity, SMs will do very well. This caution is unnecessary if we assume complete opacity: the agents will not know what population distribution they are facing.

[227] Danielson, 158-159.

Agent	UC	CC	RC	SM
UC	3	3	1	1
CC	3	2.75	2.75	2
RC	4	2.5	2.5	2
SM	4	2	2	2

Table 10.1 Danielson's Concocted Scrutiny Costs[228]

Applying these costs to our variables above, we have the following tournament play scores:

CC = 3(2)+2.75(10+5)+2(3)-2.75 = 50.5
RC = 4(2)+2.5(10+5)+2(3)-2.5 = 49
SM = 4(2)+2(10+5+3)-2 = 42
UC = 3(10+2)+(1)(3+5)-3 = 41

[228] Danielson, 159. Two insignificant modifications were made to the original table. Danielson's payoff structure in his table was as the following: being cheated = 0; mutual defection = 1; cooperation = 2; and cheating = 3. To match our payoff structure, I have altered these scores so that 0=1, 1=2, 2=3, and 3=4. Also, instead of SM, Danielson speaks of unconditional defectors (UD). UDs are not identical to SMs, in that SMs might actually cooperate if they knew that the others would cooperate if they did and SM acted first in a part-forward, part-spot contract. This is so because SMs wish to maximize, and they wouldn't maximize by defecting in such a situation. UDs, on the other hand, are hard-wired to defect whatever the situation. For this difference to be applicable, however, part of the condition is that SM *knows* what the other player will do. Such knowledge requires transparency. In our discussion, we are considering the costs of translucency, and so we shall assume UDs are SMs for all intents and purposes.

Danielson admits this much: "[G]iven differential epistemic costs, the superiority of RC over CC is weakened."[229] It is not merely "weakened," however, it shows that CC is the superior strategy. At least CC is the superior strategy given this rather arbitrary concoction of scrutiny costs.[230] It seems too important a point to let hang like this, however. Rather than a mere guess at the extent of the costs of scrutiny, we need a concise formula for working out the full ramifications and complications involved in introducing translucency conditions. What follows is my attempt at doing so.

10.2 THE MAKING OF THE TRANSLUCENCY FORMULA

The translucency formula is an addendum to the already existing transparency formula. It is simpler to do it this way, as the reader will no doubt soon appreciate. For RC, for example, every encounter with one whom he suspects to be a UC, yields a 20% chance of getting only 2 points, since if the UC is a CC or another RC, they will see his intention to defect, and thus they will defect as well. Meanwhile an SM would defect anyway. Thus, so long as an RC fails to accurately assess the other player as being a UC, he will get 2 points in that encounter. In his encounters

---

[229] Danielson, 159.

[230] Since he "concocted the outcomes,...not much hangs on them," Danielson, 159.

with persons whom he suspects to be CCs and RCs, there is a chance that they are actually SMs, thus he would only get 1 point for those encounters. If an SM encounters a player whom he suspects to be a UC but is really another SM, then rather than 4 points, the SM gets only 2 points. Thus SM loses 2 points for every encounter with an SM that the SM suspects to be a UC. Assuming the calculations for encounters under transparency have already been tabulated, we need subtract or add points according to the patterns revealed under our translucency condition. The following table is a helpful guide to the alterations needed for any situation of mistaken identity with the four dispositions so far identified.[231]

	SM	RC	CC	UC
SM SM	n/a	=	=	+2
SM RC	=	n/a	=	+2
SM CC	=	=	n/a	+2
SM UC	-2	-2	-2	n/a

SM Total score = 0

	SM	RC	C C
RC SM	n / a	=	=
RC RC	-2	n/a	=
RC CC	-2	=	n / a
RC UC	-2	-2	- 2

RC Total score = -8

---

[231] When we later introduce a fifth disposition, this table will have to be expanded.

		SM	RC	CC	UC
CC	SM	n/a	=	=	+2
CC	RC	-2	n/a	=	=
CC	CC	-2	=	n/a	=
CC	UC	-2	=	=	n/a

CC Total score = -4

		SM	RC	C C
UC	SM	n/a	=	+ 2
UC	RC	=	n/a	+ 2
UC	CC	-2	-2	n / a
UC	UC	-2	-2	=

UC Total score = 0

Table 10.2 Expected values for Translucency Conditions ( ) [232]

Whether the UC gains any points, loses any points, or remains at the same point structure he would have received under conditions of transparency will be determined by which disposition he is really encountering. For example, when an SM believes incorrectly that he is encountering a UC, but he is really encountering another SM, he does not stand to gain the 4 points SMs usually get for their encounters with UC, but instead will only get 2, since his co-player is really an SM. Thus he stands to lose 2 points, since the scoring under the transparent condition has already been calculated.

[232] In this table, the notation, " " indicates the relation "mistakenly believes he is encountering." Thus, "UC SM" reads, "a UC mistakenly believes he is encountering an SM."

Most of the scoring should be straightforwardly understandable. One aspect may need further explaining, however.[233] An RC mistakenly believing he is encountering a UC but who is really encountering a CC loses 2 points. One might suspect he should lose only 1 point, for an encounter with a CC normally gives him 3 points, which is only one less than the 4 points he gets in his encounters with UC. But we must make an assumption consistent more with transparency conditions. We assume the RC is mistaken, not the CC. But if the CC thinks the RC is going to cooperate in this encounter (as they would under transparency conditions) he would be mistaken, for the RC is going to defect, thinking he is dealing with a UC. Thus, we assume, however clumsily, that the misidentified co-players (the CC in this case) can accurately assess the *intentions* of the misinformed player, rather than the disposition. Thus, we assume the CC will be able to tell that the RC will defect in this case. Hence the CC will defect as well. It is for this reason they each

[233] Another point may also require clarification. In the case where a CC, say, believes he is encountering a UC but is really encountering an SM, the CC will lose 2 points. But there is no apparent avenue for the SM to receive those 2 points in that transaction. This is not true. Rather than SMs gaining points by this formulation, the CC and RC dispositions lose points. Thus, although I spoke originally of the SM "gaining" points under translucency, this formulation takes that into account by SMs *relative* performance compared to the performances of the other dispositions in the tournament. When our task is to determine the dominant strategy, this difference in scoring should not matter. A point differential of 2, say, still exists, whether it is because the SM competitor loses 2 points, or an SM gains 2 points. If as well as the CC losing 2 points, the SM gains 2, this means we are doubling the effects of translucency in favour of SM, and this is illegitimate.

get 2 points rather than 3 points. As 2 points are two less than 4 points, we must deduct 2 points from the RC under these and similar situations.

We see from the chart that RC stands to lose the most points under conditions of translucency. CC also loses points, but not as many. SMs and UCs lose no points whatsoever. Their scores remain identical to scores under conditions of full transparency. This should not be surprising, since there is no scrutiny cost for either UC or SM. Moreover, with the added feature of needing to distinguish CCs from UCs, which CCs need not bother with, the chart also captures the differences in scrutiny costs between CCs and RCs.

But the story is not yet done. We can not simply deduct 8 points from RCs total score under the transparency conditions (giving RC 48 points), and 4 from CC (giving CC 50) as if that is all there is to showing CC to be the superior strategy, contra Danielson. Not only do we need to take the probability of mistaken identification for every encounter into effect, we need also take into account the probability of the mistaken encounter being with *this* particular disposition. To do this more complicated step, I propose the following formula:

$$T + (\ _{i})(r\text{-}1)(D\text{-}1).$$

where T = transparency outcomes; = the translucency modification scores derived from table 10.2; D = the total number of dispositions in the pool; and r = the

probability of correctly identifying the other player's disposition under conditions of translucency. Thus, D-1 = the number of dispositions that one's co-player might actually have in case one misidentifies the other player's disposition. In our scenario, so far, there are 4 dispositions. One of those four dispositions has been accounted for already under the transparency calculations, thus the possible dispositions under a mistaken encounter is 3. The probability of being mistaken in a given encounter is represented by the notation: r-1. Our going assumption was that for every encounter we suffer 20% chance of being mistaken. Hence r-1 = .20. To understand how translucency is going to effect our players, we need to multiply the scores shown in table 10.2 with this formula. The result is:

CC Total score = $T + (_c)(r\text{-}1)(D\text{-}1) = 54 + (-4)(.2)(3)$
RC Total score = $T + (_r)(r\text{-}1)(D\text{-}1) = 56 + (-8)(.2)(3)$
SM Total score = $T + (_s)(r\text{-}1)(D\text{-}1) = 42 + (0)(.2)(3)$
UC Total score = $T + (_u)(r\text{-}1)(D\text{-}1) = 41 + (0)(.2)(3)$

Which works out to:

CC Total score = 51.6
RC Total score = 51.2
SM Total score = 42
UC Total score = 41

CC still wins out, albeit not by much. Unfortunately for CCers, the story does not quite end here either. We need also take into account population differences of

dispositions. If the population were equal among all dispositions, say 10 UC, 10 SM, 10 RC, and 10 CC, then the above translucency formula for modifying the transparency calculations would be sufficient (with appropriate modifications to the transparency figurings to take into account the different populations).[234] If there is less chance for certain mistakes to occur, as there would be if there are less of one disposition than another, a further calculation needs to be made; this:

(d/n)

where d = The number of players with a particular disposition in the pool; and n = the total number of players in the entire pool. Thus, d/n = the likelihood of an encounter with a player with a particular disposition.

We cannot simply multiply the total score above with d/n, since d/n must be calculated at the level of individual interaction. The above transparency scoring chart needs to be altered to take population distributions into effect. Assuming our populations are as originally recorded (10 CCs, 5 RCs, 3 SMs, and 2 UCs), the translucency modification chart taking d/n into account will be as shown in table 10.3 (on the following page). Note how in some cases, I used d-1/n. This is necessary because although you may not know which disposition you are actually

[234] With a population of 10 UCs, 10 SMs, 10 RCs and 10 CCs, the following scores will result: CC = 104.6 [3(10+10+10)+2(10)-3+(-4)(.2)(3)]; RC = 112.2 [4(10)+3(10+10)+2(10)-3+(-8)(.2)(3)].

dealing with, you know it is not yourself. We must subtract 1 from d when d is the number of players with the same disposition as the player under investigation.

		SM	RC	CC	UC
SM	SM	n/a	=	=	+2(2/20)
SM	RC	=	n/a	=	+2(2/20)
SM	CC	=	=	n/a	+2(2/20)
SM	UC	-2(2/20)	-2(5/20)	-2(10/20)	n/a

SM Total score = -1.2

		SM	RC	CC	UC
RC	SM	n/a	=	=	+2(2/20)
RC	RC	-2(3/20)	n/a	=	=
RC	CC	-2(3/20)	=	n/a	=
RC	UC	-2(3/20)	-2(4/20)	-2(10/20)	n/a

RC Total score = -2.2

		SM	RC	CC	UC
CC	SM	n/a	=	=	+ 2 (2 / 2 0)
CC	RC	-2(3/20)	n/a	=	=
CC	CC	-2(3/20)	=	n/a	=
CC	UC	-2(3/20)	=	=	n/a

CC Total score = -0.7

		SM	RC	CC	UC
UC	SM	n/a	=	+2(10/20)	+2(1/20)
UC	RC	=	n/a	+2(10/20)	+2(1/20)
UC	CC	-2(3/20)	-2(5/20)	n/a	=

UC	UC	-2(3/20)	-2(5/20)	=	n/a

UC Total score = +0.8

Table 10.3 Expected Values for Population Sensitive Translucency Conditions [($_i$)(d/n)]

Accounting for these population sensitive figures [($_i$)(d/n)] in our translucency formula, [(D-1)(r-1)] and attaching these to the transparency scores (T) gives this final result:

RC Total score = 56 - 2.2(3)(.2) = 54.68
CC Total score = 54 - 0.7(3)(.2) = 53.58
UC Total score = 41 + 0.8(3)(.2) = 41.48
SM Total score = 42 + 1.2(3)(.2) = 41.28

And this, after all our intense calculations, keeps RC on top of CC, contrary to Danielson's concocted figurings. But note, by hardly any margin at all. Moreover, our formula relies on the somewhat arbitrary figure that encounters under translucency conditions are 80% accurate. If we do worse than that, RC will not do as well.

The final formula for introducing the complications of translucency into game theoretic analysis is the following:

$$T + [(\ _i)(d/n)](r-1)(D-1)$$

The benefit of this formula is that it is applicable to any type and number of dispositions, any population size, and any outcome values.

10.3 OUTCOME VALUES

Gauthier and Danielson in their models use ordinal rankings. Ordinal rankings do not take into account the possibility that I may prefer mutual cooperation two times more than mutual defection, or being exploited 2 times worse than remaining at the status quo. More so than the translucency figure, however, we seem to be unable to non-arbitrarily decide a workable formula using cardinal ranking. The best we can do is hope for randomization effects to work the results out in the end. I mention this for two reasons. Game theory is constrained to work with ordinal rankings only to be accessible to all peoples. It may work out in the end on the liberal assumption that people are basically alike. Nevertheless, I think we are wise to keep this defect in the back of our minds when we assess the applicability of game theory analysis. For although there may be no grounds to this, we can work it out so that the wisest disposition is to be (or remain) an SM, so long as we adjust the outcome values sufficiently. Consider this pattern:

		A	
		play fair	cheat
B	play fair	3	10

	3	1
B	1	2
cheat	10	2

Table 10.4. Altered outcome values of the PD.

In this case, our scoring outcomes between the various dispositions (ignoring translucency) would look like this:

RC = 10(2)+3(10+5)+2(3)-3 = 68
CC = 3(10+5+2)+2(3)-3 = 54
SM = 10(2)+2(10+5+3)-2 = 54
UC = 3(10+2)+(1)(3+5)-3 = 41

True, this does not make SM the favourite disposition, but it does, at least, make it as favoured as CC. If our goal is to reduce the SMs in the world, increasing the payoff of successful cheating is not going to accomplish this.

One thing we can change, however, is the outcome values. If mutual defection equals the status quo, why not assign this a 0 rather than a 2? Adjusting the other scores to maintain the same ordinal ranking, we would then have this new scoring pattern: Being cheated upon = -1; mutual defection = 0; mutual cooperation = 1; successful cheating = 2. True, this will not alter the outcome pattern by itself. RC will still come out on top, and UC on the bottom.

RC = 2(2)+1(10+5)+0(3)-1 = 18
CC = 1(10+5+2)+0(3)-1 = 16
SM = 2(2)+0(10+5+3)-0 = 4

$$UC = 1(10+2)+(-1)(3+5)-1 = 3$$

10.4 POPULATION

Altering the population will alter the outcome. If we start with a state of nature we have SMs and no one else. How will the invasion of a CC or RC affect this? Not at all, until we have at least one other CC or RC (assuming a UC just cannot survive yet). Then these CCs and RCs will start doing better. When will an RC start outshining a CC? Only once a UC enters the picture. But UCs cannot enter the picture until the SMs have been significantly reduced in comparison to CCs. Again, this should not alter the basic claim that RCs do a bit better than CCs so long as there are UCs. But how do UCs enter the picture?

Danielson argues that a new disposition can "invade" a current population only so long as the new disposition does at least as well as the current leader.[235] So long as the entire population is comprised of CCs, then UCs can invade since they can do as well as CCs. But how likely is it the case that the entire population is comprised of CCs? One plausible story, and the one I take Danielson to be assuming, is that once CCs invade a population of SMs, they can eventually displace SMs completely. This would leave a pure population of CCs. Now, I do not

[235] Danielson, 98.

believe this story is realistic. Even if it were, however, there is a problem. UCs, under this picture, can invade only so long as SMs are displaced by CCs. RC is a superior disposition to CC only when a certain number of UCs exist in the population. But once RCs invade, then just as the SMs were dislodged, so too should the UCs be dislodged. But dislodging UC dethrones RC.

Danielson argues against the claim that UCs will simply die out on the basis that there is no necessary connection between interest and survival.[236] If there is no necessary connection between interest and survival, there is no grounds to suspect SMs will disappear after a CC invasion. But so long as there are SMs in the population, UCs cannot invade, for they cannot do as well as the best disposition (CC in this case). UC, then, is an illegitimate addition under Danielson's scheme. And depriving RC of UC fodder deprives RC of its top dog status.

Perhaps this requires further explication. If the task is seeing the survival benefit of becoming "moral," we should start off with the state of nature where we assume everyone is an SM. If everyone is an SM, everyone gets zero points. If we introduce one CC, he will still defect with defectors, and so everyone still gets zero points. If we introduce instead one UC, he will not survive. If we introduce instead an RC, again everyone will keep at zero points. The outcome distribution cannot

---

[236] Danielson, 101.

change until more than one new disposition is introduced, for example 2 CCs, or an RC and a UC. But at this stage, there is no survival value for a UC, and without UCs, there is no point to being an RC.

Let us, then, introduce 2 CCs. Without the translucency supposition, the two CCs benefit by 1 point each, which is better than the zero points for all the SMs. This slight advantage is considerably lessened once we introduce the translucency clause. Our translucency formula, recall is:

$$T + [_{i}(d/n)](r-1)(D-1).$$

Assuming there are 10 SMs and 2 CCs in our population, for each of CC's encounters with whom he suspects to be another CC, CC stands to lose 2 points. $_{c}$ = -2. r-1 is still .2. There is a 20% chance of being mistaken in any encounter. D is the number of dispositions in the tournament. Here there are only two, and thus if an encounter is not with a CC, it must be with an SM. Since in this case there are 10 SMs out of a total of 12 players in the tournament, for a mistaken encounter to really be with an SM (d/n) = .83. The formula takes into account the probability that among all players in the tournament, the mistake is going to cost in this encounter. In this case, the price of translucency is going to cost CC 2(.2)(1)(.83) = 0.3. No other encounters will make a difference whether it is under conditions of

translucency or transparency.[237] Given our hypothesized population, the outcome looks like this:

CCs = {1(2)+0(10)-1} - {2(.2)(1)(10/12)} = .17
SMs = {0(2+10)-0} - {0(.2)(1)(2/12)} = 0

CCs do better, but almost insignificantly so. It would not clearly be irrational for the two CCs to switch back to being SMs. CCs start clearly doing better only so long as there are a sufficient number of them. This fact, unfortunately, introduces an assurance problem. Ignoring the complication in getting enough CCs, let us see how many CCs in relation to the population of SMs there would need to be. Let us try half. Hence think of 10 SMs and 5 CCs. The outcome is the following:

CCs = {1(5)+0(10)-1}-{2(.2)(1)(10/15)} = 3.7
SMs = {0(5+10)-0}-{0(.2)(1)(5/15)} = 0

CC is looking to be more rational than SM the more CCs there are relative to SMs. If the population of SMs are equalled by CCs, say 10 each, the following scoring occurs:

CCs = {1(10)+0(10)-1}-{2(.2)(1)(10/20)} = 8.8

[237] For example, when SM believes he is encountering another SM, but is actually dealing with a CC, no difference in score is to be expected because we assume the CC will see that SM plans to defect and thus CC will defect as well.

SMs = {0(10+10)-0}-{0(.2)(1)(10/20)} = 0

The advantage in being a CC grows exponentially the more CCs there are relative to the SM population. Minimally CCs do better than SMs so long as there is there is at least 2 CCs and any number of SMs. There is thus no assurance problem in getting more CCs to join once there are at least two. Is there an assurance problem for those first two CCs? If the player I thought was going to be a CC, changes his mind, we are assuming an 80% chance of the other CC responding accordingly, i.e. defecting. Is it worth the risk? The danger is this, for a single CC in a population of SMs:

CCs = {1(1)+0(10)-1}-{2(.2)(1)(10/11)} = -0.4
SMs = {0(1+10)-0}-{0(.2)(1)(1/11)} = 0

So long as .04 is a significant loss to players, it would appear there is an assurance problem to get the CC disposition to enter the game. Once the population consists of any number of SMs and any number greater than 2 of CCs, it is clearly rational to become a CC, and hence there is no assurance problem. But there does seem to be an assurance problem for the first two CCs, given the condition of translucency. It may well be wiser to remain an SM and await the arrival of CCs before you change your stripes. But so long as everyone is doing this, we are back in the PD.

What about the other dispositions? We know that RCs do not do better than CCs so long as there are no UCs. But what can impel anyone to become a UC? Or, to put the case differently, how many CCs relative to SMs need there be before it is safe for some UCs to appear? In one sense this question seems moot if our task is to see whether it is rational to adopt "moral" dispositions over immoral ones as far as evolutionary survival is concerned. We already have our answer: *Yes, so long as you can convince one other to do so as well.*

Below is the situation with one lone UC trying to enter a population of 10 SMs and 10 CCs. First, we need the score differentiation table.

	CC	SM	UC
CC CC	n/a	-2(10/21)	=
CC SM	=	n/a	+2(1/21)
CC UC	=	-2(10/21)	n/a

	CC	SM	UC
SM CC	n/a	=	+2(1/21)
SM SM	=	n/a	+2(1/21)
SM UC	-2(10/21)	-2(9/21)	n/a

	CC	SM	UC
UC CC	n/a	=	=
UC SM	+2(10/21)	n/a	+2(0/21)
UC UC	=	-2(10/21)	n/a

Table 10.5. ($_i$)(d/n) for a population consisting of 10 CC, 10 SM, and 1 UC.

The resulting outcome is this:

CCs = {1(10+1)+0(10)-1} - = 10
SMs = {2(1)+0(10+10)-0} - = 2
UCs = {1(10+1)+-1(10)-1} - = 0

Clearly UCs cannot invade individually. But like the CCs, can they do so in groves? Below is the situation with ten UCs trying to enter a population of 10 SMs and 10 CCs.

SMs = {2(10)+0(10+10)-0} = 20
CCs = {1(10+10)+0(10)-1} = 19
UCs = {1(10+10)+-1(10)-1} = 9

The invasion of UCs places SMs back on top. This is no good for so long as there is a significant difference between 19 and 20, this may entice a number of CCs to switch back to being SMs. We would have a cycle where SMs become the dominant group, wiping out UCs in the process, at which point it becomes rational for a group of CCs to reenter the picture. This scenario, if true, reveals an instability of dispositions. People will change their behaviour patterns between CC and SM, much in the way Hume figured would happen under differing circumstances. Danielson's RC may in reality be indecisive CCs at a time in history where SMs and CCs scores are evenly (or relatively evenly) split.

At the minimal level where CCs start doing better than SMs, no one can afford to be a UC. Not only that, CCs have reason to prevent UCs from invading since they do worse (10.8 compared to 11) while SMs benefit (10 rather than 6). At what point can someone rationally choose a UC disposition over a CC or SM disposition? We know that if the entire population were CCs, then one might as well be a UC. A UC will not do worse or better than a CC in a population comprised solely of CCs and UCs.[238] What if CCs outnumber SMs 2:1? Let us assume there are 20 CCs, 10 SMs, and 1 UC. We would have the following outcome:

CCs = {1(20+10)+0(10)-1} = 29
SMs = {2(10)+0(20+10)-0} = 20
UCs = {1(20+10)+-1(10)-1} = 19

The problem is that UCs do better only so long as there are enough of them. But the more UCs there are, the better the SM population does as well.

A lone UC does worse than either an SM or a CC, but he does considerably better than UCs do in the previous scenario. It appears UCs do better the more UCs there are, just as with CCs. The problem, though, is that SMs become the dominant strategy the more UCs there are. This defeats the purpose of becoming a CC. CC, then, would need to discourage UCs, and a good way of doing that is by

[238] In fact, UCs may do better given the low computation costs of UC compared with CC (see ahead to my discussion of computation costs, chapter 10, section 6).

becoming an RC. This should deplete the UC population, thus diminishing the gains for SMs, thus depleting the SM population as well.

10.5 KING BREAKERS

Not only can we reduce the rationality of RC by ridding UCs, we could also imagine a new disposition entering the picture. Let us imagine that someone is bothered by RC's superiority on the basis of RC's shafting innocent UC. He is so bothered by this, he considers boycotting games with RC. Otherwise he is a CC at heart, will cooperate with other cooperators, except those who defect against those who cooperate with other cooperators. Let us call this disposition "Principled Cooperators" or PC.[239] PC will not do as well as the top dog, but his presence, especially if there are enough of likeminded PCs, can clearly affect RC. So long as there are any RCs, PCs and RCs will counteract each other's success. I assume that PC does not take advantage of RC. RC, we shall imagine, can recognize PC and fail to cooperate with them. Since, PC and RC counteract each other in this way, CC will rise to the top niche, ousting Danielson's favourite RC. Assuming there are as many PCs as RCs (5, in our example) and the remaining numbers are the same (10 CC, 3 SM, and 2 UC), the following scoring will be this:

---

[239] Danielson calls this disposition "Unconditional Cooperator Protectors" (UCP), in Danielson, 118.

CC = 3(10+5+5+2)+2(3)-3 = 69
RC = 4(2)+3(10+5)+2(3+5)-3 = 66
PC = 3(10+5+2)+2(3+5)-3 = 64
UC = 3(10+5+2)+(1)(3+5)-3 = 56
SM = 4(2)+2(10+5+5+3)-2 = 52

With the introduction of PC, CC becomes once again the dominant strategy. Danielson makes a worthwhile objection against this sort of addition, however. Any superior position given the available population can be targeted by some further invention. King-breakers, as he calls them, then will be disallowed. For artificial morality this is a necessary requirement, I think. For our goal of tying game theory to the real world, however, we have a different restriction: namely what sort of dispositions are out in the real world. My principle cooperators are problematic for it does not seem individually rational to be one. The need to implement such an RC breaker seems to merely enhance some people's suspicions that the rational move is not necessarily the moral move. That a PC is moral, may not be in question. That it is rational is.

But Danielson's restriction on introducing dispositions gives the PC an in. That is to say, so long as we do not start with CCs, but have PCs invade at the outset, a PC invasion in a population of SMs is rational. The outcome will be exactly the same as CCs. UCs can enter into the picture as well (or as badly) as they can with CCs. But now an invasion of RCs cannot dislodge the current top dog, PC, and

thus RCs would be disallowed in the game according to Danielson's rule. Taking our original population and size, and substituting PC for CC, the outcome would be this:

PC = 3(10+2)+2(3+5)-3 = 49
RC = 4(2)+3(5)+2(3+10)-3 = 46
SM = 4(2)+2(10+5+3)-2 = 42
UC = 3(10+2)+(1)(3+5)-3 = 41

What this reveals is that PC, not RC or CC, is the dominant rational strategy. Being that it is also the dominant *moral* strategy, this is good news for those of us who wish to show that morality is the rational choice.

I do not think this objection is very telling, mind you. What it reveals is not that Danielson's RC is not as rationally superior as he makes out, but that it is rationally superior given the parameters of the opening populations in the game. We need next wonder whether we should be content with this opening position. What criteria do we have to evaluate it? Danielson claims only validity, and not soundness. But how do we extend the analysis to soundness when there appears to be an infinite number of possible starting points and possible first invaders? A solution is to evaluate the plausible opening dispositions according to some measure of verisimilitude. Game theory is not meant to be historically accurate, however, and so it is difficult to judge precisely what this means. I am content, at any rate, to begin with the SM disposition. After all, if it is rational to construct a

moral society from this starting position, we have shown a lot. But why should we have CC invade rather than PC? One possible answer is that PC makes no sense without RC. Otherwise, we could simply have RC invade. So we limit our second entry to the sole disposition that is minimally designed to do better. CC fits that bill. UC does not, since although it is a simpler design than CC, it cannot survive in a population of SMs. So we have two criteria for entrance taken from biology: survival and simplicity. RC and PC could survive as second entrants, but have complications that at the second entry stage are unnecessary, and hence biologically unlikely.[240] After SM and CC are on the stage, again neither RC nor PC will evolve, for RC requires UC in order for it to be rational to develop the further complication, and PC requires RC for its evolution. But here we seem to run to a standstill, as noted earlier in this section. UC cannot do as well as CC and thus should not be allowed to enter the game. And lacking UC, we lack the evolutionary niche appropriate to RC. We can only allow RC to enter so long as we loosen slightly our criteria of third level entrants to allow UCs to enter. But once we've loosened our criteria to this extent, then there can be no objection to bringing in PC.

---

[240] Some may contest that what is evolutionally possible is whatever is computationally possible. So long as evolution proceeds through both survival and random mutation, the RC and PC distributions are possible at the second entrance level. To rule them out of hand at the second level entrance stage is to erroneously regard evolutionary processes in the Lamarkian fashion rather than in the modern Darwinian sense. My point still stands, I believe, for I focus on probability.

Consider the following loose criterion for entrance:

C1: *A disposition can successfully invade a population so long as it does at least as well off as the poorest members of that population.*

The rationale here is that if these members are surviving, however poorly, then so too can others at their subsistence level. A reasonable complaint against this in terms of realism may be that although the poorest members of a population are currently surviving with scarce resources, they may no longer be able to sustain themselves within a larger population. We shall therefore tighten our loose criterion for entry to this:

C2: *A disposition can successfully invade a population so long as it does better than the poorest members of that population.*

For us to allow UC into the game, UC must do better than the worst disposition. Can UCs enter a population of SMs and CCs under this entrance test? If there are 5 CCs, 5 SMs, and 2 UCs, the scoring under conditions of transparency, is CC = 28, SM = 26, and UC = 23. Since UCs do not fare better than SMs, they could not enter this population. Is there a population distribution of CCs, SMs and UCs in which they could enter? Yes. Imagining that there are 10 CCs, 2 SMs and 2 UCs, UC can score better than SMs. Hence, we know that under certain population conditions,

UC does better than SM, and so can be allowed to enter the tournaments. Now RC storms in.

The point is that only by lowering our standards of entrance can Danielson's RC enter the game. There is no population arrangement with SMs, CCs, and UCs where UCs do better than the other dispositions. Lowering our standard of entry, however, also lets in PC, since PC does better than the lowest disposition in the pool. Once PC is allowed to invade, this upsets the apple cart of RC. CC becomes the dominant strategy, not RC.

10.6 COMPUTATIONAL COSTS

Computational costs will differ for the different game strategies. Recall Danielson's introduction of concocted scrutiny costs.[241] This is ambiguous. There are costs of making mistakes in one's detecting other disposition types and there are costs of simple processing.[242] The first sort of costs come under the rubric of translucency costs, already discussed. The second sort of costs are computation costs that would be incurred even under conditions of transparency. Danielson did not need

---

[241] Danielson, 159.

[242] There are other costs as well. Danielson notes costs of exposure. Exposing one's intentions, necessary for all but the SM disposition, runs the risk of not only being then exploited, but also being thrown questions which will put them in endless loops (see Danielson, 156). But this is not the computation cost I am concerned with here.

to make this distinction since he was dealing with a virtual world where computation is free and fast. People are not so lucky.

UC and SM require no added processing at all, and hence no difference in scores is to be expected. CC needs to distinguish between cooperators and non-cooperators, and thus require some computation costs greater than UC or SM. RCs, meanwhile require further computation than CC. Not only do they need to distinguish cooperators from non-cooperators, and thus incur the same computation costs as CC, they also need to distinguish conditional cooperators from unconditional cooperators. Thus RCs should be expected to suffer the most computation costs, CC next, and UC and SM none at all.[243]

We need to decide two things: what the costs are that we should assign and whether these costs differ according to the different dyads being assessed. Concerning the first, if we incorporate Danielson's figurings from table 10.1, CC incurs a computation cost of 0.25, while RC suffers a cost twice that. One might object that this cost has already been accounted for by our discussion of translucency. Translucency costs are incurred from mistaken encounters, and not

[243] At least given the algorithms of CC and RC given by Danielson. Paul Viminitz has indicated to me that this does not mean that all possible algorithms for CC and RC will invariably show RC's algorithms to be more complex than CC's. We can imagine, for example, CC executing exactly RC's algorithm, and then adding a new algorithm to count any disposition to cooperate at all as meriting cooperation now. In such a case, however redundant it is for CC, RC would have less computation than CC.

otherwise. Computation costs, on the other hand are to be added for each encounter, regardless of which disposition one is encountering, and regardless of correct or mistaken identity. Danielson's explanation for why RC should have a greater cost than CC is because "RC pays even higher costs than CC, since it must test agents *counter-factually*: asking not merely `[would] you cooperate', but `would you still cooperate were I to defect'?"[244] This is not the rationale for the greater costs incurred by conditions of *translucency*. Thus we are left to wonder whether Danielson's *scrutiny* costs are not my *computational* costs. This would seem to follow given that the extra computation involved is because of the requisite *scrutiny* needed by CC and RC players. Moreover, as I have argued, my formula for translucency does not yield the same result as Danielson's scrutiny costs in table 10.1. Nevertheless, I am not convinced Danielson's concoction of scrutiny costs can be incorporated as our computational costs. For one, Danielson also remarks that "responsive agents, because they pay costs of exposure and scrutiny, will do relatively less well as we move from perfectly transparent conditions."[245] whereas my computational costs should be in effect even in conditions of transparency; no difference in computation costs is to be expected between conditions of

[244] Danielson, 158-159.

[245] Danielson, 157.

transparency and translucency or even opacity. Those differences are matters of the effects of error, not computation. Another problem with assuming the figures in table 10.1 are computation costs is that we would expect CC and RC to suffer computation costs in *every* encounter, not just those with CCs and RCs. This follows because CC still must distinguish between SM and UC in order to know how to act in this encounter. SM and UC do not. This is likewise the case with RC. Not only must RC be able to distinguish SM from any other disposition to avoid being cheated, RC must also distinguish between CC or RC and UC. And this is the case whether or not their detection is 100% accurate.

Although we are not yet sure how high these computation costs are, we can see that it may be deemed too much an effort to be an RC for some people -- even if after the process they are always right in their detection. The computational burden may be too great. Especially since the overall score of CC and RC without incorporating computational costs are not that far apart.

Our problem is assigning some value to these costs. We can assume that encounters will differ in time constraints. Some games require immediate action, others not. When time constraints are not a factor, greater computation is available at relatively little cost. When time is a factor, however, individuals may not be able to afford the extra computation necessary to be an RC. It is unlikely, given the variability of time factors in encounters, that computation costs are as high as the

computation costs Danielson projects in his scrutiny cost table (table 10.1). In our scoring system, table 10.1 indicates that RC incurs a cost that is 17% of his total intake with every CC player.[246] Surely this overstates the case. Let us assume that a reasonable cost per encounter for CC is .1, and .2 for each encounter for RC.[247] Because these costs are not *differential* costs, as is the case for the costs of translucency, we can apply this added cost quite simply.

$$T_i - C_i(n-1)$$

where T_i is the Translucency score for the given disposition, C_i is the computation cost assigned to the given disposition (.1 for CC and .2 for RC and 0 for UC and SM), and n is the total number of players in the tournament. Our scoring outcome, then, looks like this:

CC's score = T_c - .1(n-1)
RC's score = T_r - .2(n-1)
UC's score = T_u
SM's score = T_s

[246] It is worse with the original scoring that Danielson uses. If RC normally gets 2 points with his encounter with CC, and loses 0.5 points in the encounter, computation costs are 25% of the profit.

[247] Where CC normally would earn 3 points, he now earns 2.9 points -- a 3% computation tax. Instead of RC earning 3 points prior to the computational costs, he now earns 2.8 points -- a 7% tax.

The result, with our standard population of 10 CC, 5 RC, 3 SM and 2 UC, and assuming conditions of translucency, is the following:

CC's score = 53.6 - .1(19) = 51.7
RC's score = 54.7 - .2(19) = 50.9
UC's score = 41.5
SM's score = 41.3

CC, albeit not by a lot, is revealed to be the dominant strategy.

I will not stand firm on this quibble. The costs of computation may not be as large as I make out. I have chosen a small fraction as the cost, but perhaps it is even less than that. Moreover, some people may be better suited to greater computation, meaning that computational costs are not the same for everyone. Perhaps for those more equipped, RC is the dominant strategy. My point is to complain that we cannot be as confident in the results of game theory as previous game theoreticians have been.

10.7 CONCLUSION

Notwithstanding this last comment, there is a general trend from game theoretic approaches to morality that is encouraging: dispositions that approximate morality also approximate rationality. I temper my conclusion far more than Danielson or Gauthier do because there are many variables that cannot be accommodated across the board. Which disposition is most rational depends on which dispositions

are already in the population; which dispositions are allowed to enter the population; what the population size is for each disposition in the tournament; what outcome values we assign for each encounter; whether players are transparent or translucent, and if they are translucent, to what extent are they accurately readable; and what particular costs players incur, if anything, for the more complicated calculations necessary for CC, RC, and my king-breaker PC. We can be confident in showing validity, as Danielson recognized, but given the enormity of variables in which we are dealing when we try to tie game-theoretic findings to real world moral problems, we are much less confident in claiming anything close to soundness.

If there is any one point important to these findings, however, it is to see the benefit of changing our stripes from straightforward maximization to one or another form of constrained maximization.

PART IV

THE SOCIAL & POLITICAL IMPLICATIONS

OF OCCURRENT CONTRACTARIANISM

CHAPTER 11

THE STATE AND PROPERTY

11.1 HOBBES

By far the most well known objection against Hobbes's account of contractarianism focuses on the inadequacy of an absolute sovereign. That we must *unconditionally* submit to the sovereign strikes many not only as unpalatable but unfounded.

For Hobbes, one should obey the moral codes of the sovereign because one will inevitably suffer otherwise. The sovereign did not create the State of Nature as a threat. It is not a Mafia ethics that Hobbes advocated. Rather, individual suffering is the natural condition and the sovereign is the solution. Without the sovereign, one would be left to defend oneself continually against the threat of men. One immediate problem is that if people are as vicious as Hobbes makes out, why should the sovereign be any better? In fact, we have every reason to assume life under the sovereign would be far worse than in the State of Nature. The sovereign's character would be the same as any man and his power would be immensely superior. We have, in fact, elected the Thrasy-machean brute to rule us. It is for this reason that Locke sarcastically remarks

that Hobbes evidently conceives men to be so "foolish that they take care to avoid what Mischiefs may be done them by *Pole-cats* or *foxes*, but are content, nay think it Safety, to be devoured by Lions."[248] An obvious answer to this "who-guards-the-guards" problem is to show that the ruler needs to rule well lest he is dethroned violently; but Hobbes presents the case that we abandon *all* power to him, and that it would be immoral to contest the ruler in anything. Hobbes is adamantly against insurrection of any sort.[249] Another problem with Hobbes's belief that we must abandon all power to the sovereign is that it is inconsistent with the model of human psychology that Hobbes accepts. Survival, for Hobbes, is above all a duty to oneself. Thus if the sovereign threatens your own life, retaliation or avoidance seems called for. To claim either retaliation or even avoidance as being immoral is to abandon, not secure, the link between morality and individual rationality. As if these two complaints were not enough, a further

---

[248] John Locke, *The Second Treatise on Civil Government*, Buffalo: Prometheus Books, 1986 [Originally published as *Two Treatises of Government,* 1690] VII, 93 (53).

[249] "They that are subjects to a Monarch, cannot without his leave cast off monarchy, and return to the confusion of a disunited Multitude, nor transferre their Person from him that beareth it, to another Man, or other assembly of men" (Thomas Hobbes, *The Leviathan*, Buffalo, New York: Prometheus Books, 1988, xviii, 90-91). See also *Leviathan*, xv (76). Some feel there are historical-sociological reasons for Hobbes's insistence on the absolute power of a sovereign, rather than purely philosophical ones. There was, accordingly, endless insurrection during Hobbes's time, and an absolute sovereign would clearly put a stop to that. See, for example, Jean Hampton (*Hobbes and the Social Contract Tradition*, Cambridge: Cambridge University Press, 1986, 110). According to A.E. Taylor, *The Leviathan* "was written in a time of revolution and unsettlement as a persuasive for cessation from fruitless civil strife" (A.E. Taylor, "The Ethical Doctrines of Hobbes," in Baumrin (ed.) [Originally published in *Philosophy*, XIII, 1938, 406-424] 36).

problem Hobbes faces is this: his claim is that people can not get along well without a sovereign to oversee their conflicts and settle disputes. Nevertheless, the claim is that we come to recognize the need for this sovereign, and we all voluntarily decide to give up our "rights." But how can we come to such an amicable agreement *prior* to the sovereign's rule? If we can come to agreement on that issue, then what need do we have of the sovereign for other agreements? On the other hand, if we require the sovereign to settle all our disputes and oversee all agreements, we could never get out of the State of Nature for there is no sovereign yet to oversee the election of the sovereign. As Hampton remarks, "Hobbes's account of conflict [in the State of Nature] seems to generate sufficient strife to make the institution of the sovereign necessary, but too much strife to make that institution possible."[250]

Locke can solve these problems because he does not envision the people relinquishing all of their rights to the sovereign. People merely lend their power. They are able to do this since Locke's conception of the State of Nature is more amicable than Hobbes's picture.

Part of the reason for Hobbes's insistence on absolute obedience to the sovereign was his belief that anything less would revert to utter chaos. What sort

---

[250] Hampton, 136.

of ruler would there be if the people reserved the right to govern him themselves?[251] Hobbes's belief is that a contract between a ruler and the people limiting the sovereign's power will require a judge to decide if there was or was not a breach. The judge can not have absolute power, however, and is thus subordinate either to the sovereign himself, in which case the sovereign has absolute power, or someone above the sovereign has power over the judge; but then this individual is the absolute sovereign. The point is: since the judge cannot have full power, he must be limited by a greater power. Just as Aquinas derived a being called God, so Hobbes derives the necessity of an absolute sovereign. Anything less is "a recipe for war."[252] This "infinite regress" argument works also for the problem of interpreting rules. People, being what they are, will interpret rules for their own advantage. If interpretation of rules is left up to individuals, this will be as having no rules at all. A final arbiter is needed if men wish to successfully escape the State of Nature.[253] Only an absolute ruler, acting

---

[251] Kant agreed with Hobbes on this point. "The permissibility of rebellion would "make all lawful constitution insecure and produce a state of complete lawlessness" (Immanuel Kant, "On the Common saying: 'This May be True in Theory, But it does not Apply in Practice'," in *Kant's Political Writings*, Hans Reis (ed.) Cambridge: Cambridge University Press, 1970 [1793], 82).

[252] Hampton, 103.

[253] Hobbes, *Leviathan*, xviii (91).

as final arbiter and final judge, can prevent the return to the State of Nature. Since prevention of the State of Nature, or war, is the sole purpose of erecting a sovereign, the sovereign must be absolute.

Hobbes's mistake lies in assuming that because we may need a judge to judge between the sovereign and the people, that the judge has powers supreme in all respects to the sovereign.[254] He neglects the possibility that the judge's power may be clearly delimited. That Hobbes requires someone with final authority over all decisions does not necessitate by itself that authority resides in one individual. If human nature is really such that we need an arbiter for every interaction, then we can not settle on an arbiter; nor, for that matter, on the arbiter's decision. If we require arbiters for only certain situations, then there is no reason to grant that individual power over more than that decision. Hobbes's argument for the *absoluteness* of a sovereign does not follow.[255]

11.2 LOCKE

Given the incoherence of Hobbes's contention that we must abandon all power to the sovereign, Locke's solution should be a relief to contractarians. It is not,

---

[254] Hobbes, *Leviathan*, xviii (91).

[255] Hampton makes an even stronger claim against the logic of the Leviathan (Hampton, 206).

however, since he succeeds only by abandoning contractarian grounds in the process. Locke's solution, in brief, is that people do not "surrender" their power to the sovereign; they lend it. For Locke, a ruler should not have the ability to interfere in people's lives. For Locke, "the supreme power cannot take from any man any part of his property without his own consent.... Men have such a right to the goods...that nobody hath a right to take them."[256] As people have this property by "right," it is *unjust* to take it away; not merely problematic to do so. We need to assess why the sovereign for Locke is limited in this way, and whether we can arrive at the same conclusion through other grounds more conducive to contractarianism.

Locke is a contractarian in many respects; but not all. As with Hobbes (and Hume and Rousseau), he believed people were roughly equal in strength and mental ability.[257] Also with Hobbes, he believed there was a close connection between self-interest and the laws of nature.[258] But these two starting points take Locke in a different direction from Hobbes. Within Locke's State of Nature there

---

[256] Locke, 2T, XI, 138 (77).

[257] Locke, 2T, VI, 54 (32).

[258] Locke, 2T, III, 19 (15).

will be frequent cooperation and successful commerce.[259] Rough equality will deter widespread violence, not inflame it. Singleton combat will only have a fifty percent chance of success given the assumption of rough equality. Thus it would be foolhardy to attack, especially since it would always be prudent to defend. Defenders have already lost the status quo (peace), so they have every good reason to fight. Given this, the incentive structure is biased toward defence, not attack. Thus there should be less violence in the State of Nature than Hobbes conceived.[260] Hobbes believed this fact would not deter attackers because attackers would tend to band together, thus increasing their chances of success. This is true, but true only so long as defenders did not band together. And there is nothing in Hobbes's argument to make us believe the contrary. It is conceivable, then, under Locke's account of the State of Nature as it should be in Hobbes, that people can conspire to hand over power to a sovereign to help settle disputes and set up a police force. The Lockean State of Nature, mind you, is not so amicable that there is no need for a sovereign. Although

---

[259] Locke, 2T, II, 14-15 (14).

[260] An early argument to this effect can be found in Samuel Pufendorf, *The Law of Nature and Nations*, C.H. and W.A. Oldfather (trans.) Oxford: Clarendon Press, 1934 [1688], 170. For a more recent argument, see Anthony de Jasay, *Social Contract, Free Ride: A Study of the Public Goods Problem*, Oxford: Oxford University Press, 1990, 41-42.

cooperation can and does exist in the Lockean State of Nature, it is a limited sort.

> [F]or all being kings as much as he, every man his equal, and the greater part no strict observers of equity and justice, the enjoyment of the property he has in this state [the State of Nature] is very unsafe, very insecure.[261]

Where Locke differs from Hobbes the most is in Locke's adamant denial of the need and justness of a sovereign's being absolute. Granting absolute power to a sovereign will merely create a Thrasymachean brute and this will be worse for the subjects than the State of Nature. If people are roughly equal in strength and ability, at least through confederacy, then nothing prevents a gang of people from banding together and usurping the tyrannical sovereign. This would seem to follow given Hobbes's characterization of human psychology. Hobbes could not accept this, for he thought it was an admission of never being able to escape from the State of Nature. As discussed earlier, Hobbes believed that the sole alternative to Absolute sovereignty was the State of Nature, since any constraints on the Sovereign's power would inevitably, and perhaps quickly, lead to absolute loss of control. This implication does not arise for Locke (nor should it have for Hobbes). Rather, the very possibility of confederacy against a

---

[261] Locke, 2T, IX, 123 (69).

poor sovereign is precisely what will keep the sovereign in check. The "best fence against Rebellion" is to rule well.[262] Peace is best served by granting power to the sovereign *on condition* that he rules well. This is assessed by seeing whether people live in peace or not. The ruler, then, is seen as an agent of the people,[263] conscripted through contractual agreements.[264] This latter part was found offensive to Hume, since historically sovereignty was assumed merely through conquest. I shall take that complaint up later. For now, the point is that given contractarian premises which Hobbes himself endorsed, the sovereign is argued to be limited in power.

Following Locke so far, we can successfully defend the limited role of the sovereign on contractarian grounds. It is not thought necessary, let alone essential, that contractarianism implies anything about absolute sovereignty. Contractarianism is the claim that people independently will do better in their pursuit of their good in a stable, peaceful society than in the State of Nature. So long as this is secured by a limited sovereign, then it is rational for the people to obey this sovereign so long as it remains in their considered self-interest to do

---

[262] Locke, 2T, XIX, 226 (121).

[263] Hampton, 123.

[264] Locke, 2T, X, 122 (69). See also 2T, VIII, 95 (54).

so. Apart from Hobbes's belief in the Thomistic infinite regress argument, there is no reason Hobbes himself should have denied this.[265]

Hobbes believed that it would not be in our interest to dispel even a tyrannical sovereign for the only alternative would be the State of Nature. Although Hobbes's State of Nature is more dire than Locke's, it is still hard to imagine that it would be worse than wholesale genocide, if such was the practice of one's sovereign. Hobbes happened to think the sovereign would not behave this way,[266] but he cannot argue that without also showing that it is in the interest of the ruler to rule benevolently. It cannot be in the interest of the ruler to rule "well" unless there is the possibility of being deposed, one would think. Even granting Hobbes the scenario that sovereigns will tend to rule well, or at least that the State of Nature is worse than any state under a sovereign, it does not follow that the people can not do better if they dispose of him. There is another option, surely, than a retreat to the State of Nature. The people could elect *another* sovereign. Moreover, if he benefits from being a sovereign, he will want

---

[265] In fact, he seems to recognize at least some limitation to a sovereign's power when he condones certain rebel activity in a commonwealth. Hobbes admits that one certainly has the liberty to defend one's life if one is expecting death at the hands of the sovereign (Hobbes, *Leviathan*, xxi, 115).

[266] The Sovereign is "carefull in his politique Person to procure the common interest" (Hobbes, *Leviathan*, xix, 98).

to retain that title. If he is aware of the possibility of dethronement, there is greater chance of his ruling better. And it is because the method of *lending* power, as Locke conceived it, rather than bequeathing it, is conducive to good rule that it would be in the people's interest to advocate it.[267]

11.3 LIMITED SOVEREIGNTY

Locke's account shows up more fully the workings of "agreement" and "contract" in contractarian thought. In contrast, absolute obedience to a ruler in the Hobbesian sense is merely a slave/master relation. Locke's social contract theory follows more closely the way agreements in commerce operate. So long as I have an interest in agreeing to specific terms with you, and you have an in interest in agreeing to the terms with me, then we both have an interest in abiding by those terms, lest the other pull out. But pulling out of the deal without extenuating circumstances would be irrational assuming the interests all participants have in the deal. Bilateral reneging returns the players to the State of Nature. So far, we can follow Locke's emendation of Hobbesian contractarianism. The problem begins, I believe, when Locke takes his argument one step farther and by doing so leaves contractarian grounds.

---

[267] Hampton makes this point, 247-248.

For Locke, absolute power is not only unnecessary to achieve peace, it is "unjust."[268] Unlike Hobbes's account, no man in a State of Nature has the liberty to harm another.

> The State of Nature has a law of Nature to govern it, which obliges every one, and reason, which is that law, teaches all mankind who will but consult it, that being all equal and independent, no one ought to harm another in his life, health, liberty or possessions....[269]

There are no contractual grounds for making this claim hold within the State of Nature. Not surprisingly, the grounding for it is theological. The quote above continues thus:

> ...for men being all the workmanship of one omnipotent and infinitely wise Maker; all the servants of one sovereign Master, sent into the world by His order and about His business; they are His property....[270]

Hobbes, recall, was looking to derive morality from an amoral state. If Locke was also trying to carry out this pursuit, he cheats by slipping into the State of Nature

---

[268] For example, "the supreme power cannot take from any man any part of his property without his own consent" (Locke, 2T, XI, 138, 77). Likewise, the Sovereign's power is constrained by "the law of God and Nature" (Locke, 2T, XI, 142, 79).

[269] Locke, 2T, II, vi (9).

[270] Locke, 2T, II, vi, (9-10).

a priori moral codes. The question we need to address is whether we can derive the Lockean claim-rights that restrict absolute sovereignty on Hobbesian grounds. We have seen that a limited sovereignty should serve Hobbes's purpose. And if that serves the purpose, the people have prudential reasons to not relinquish their liberty any further than the limited restrictions. But to claim there is no "right" to do so is what a contractarian seems unable to aver.

Before proceeding, let me review the arguments of this chapter. Hobbes's position on absolute sovereignty is incoherent. What he could consistently defend, however, is what Hampton calls the "fallback position."[271] This, in brief, is Locke's claim that people do not "surrender" their power to the sovereign; rather they lend it. For Hobbesians to agree to this rendition, they must give up Hobbes's "infinite regress" argument. This was the claim that any ruler who is subject to checks by the people will be no ruler at all.[272] Even if contractarians adopt the "fallback position" (and doing so naturally follows from the Hobbesian psychological premises), it is not yet clear that they can make the same libertarian restrictions on the ruler that Locke makes.[273] For Hobbes, the ruler

---

[271] Hampton, 220.

[272] Hampton, 98-105; 224. Hobbes, *Leviathan*, xviii-xx.

[273] "An agency-agreement *ought* to contain provisions against interference...in religious affairs, economic transactions, and private activities" (Hampton, 250, my emphasis).

must have the power to command his subjects in *any* area of their lives. This is wrong. People are able to escape the State of Nature without this need for an absolute sovereign. For Locke, the ruler *must* be restricted in his range of power. This too is wrong unless we qualify it by making it merely a prudential claim. If we can escape the State of Nature with either a limited or an absolute sovereign, we will quite likely do better, prudentially considered, with a limited sovereign. Thus we *should* restrict the sovereign's power. But making the stronger claim that an absolute sovereign exceeds the bounds of "Rightness" is beyond the ability of contractarians to stipulate. Locke can not ground his right-based position on contractarian grounds.

Hampton admits that on the basis of the agency-relation in the fallback position, there are clear limits as to how far the government can go. The rulers fear being dethroned if they step on too many toes.[274] Once we adopt the fallback agency-relation position the power of the ruler is considerably constrained. The ruler has only the power so long as the people choose to obey his punishment commands.[275] Nevertheless, Hampton contends, this power is too much.[276] I do

---

[274] Hampton, 252.

[275] Hampton, 250.

[276] Hampton, 251.

not think this follows. Recall, the agency-relation between the sovereign and the people is such that (1) people can agree on how well the ruler is performing his task, (2) that it is in the ruler's interest to perform that task well to avoid being deposed by the people, and (3) that it is in the interests of the people to depose the ruler if he fails to satisfy the people's desires, which Hobbes thought was predominantly the securing of peace. Concerning the viability of (3), the State of Nature is not the only alternative if we dispose of this ruler; another ruler more conducive to our needs and desires is what we will be looking for. Given this, the limitations Locke speaks of will in fact be the limitations of which a Hobbesian fallback ruler and his subjects will be aware. But if both Locke's position and Hobbes's fallback position amounts to the same sort of limitations on a ruler, they arrive at these considerations in different ways. In neat, Hobbes by self-interest solely; Locke by some other route.

11.4 JUSTIFICATION FOR THE LIMITED STATE

Locke maintains that no ruler has a right to interfere with the people's liberties or property. The fallback Hobbes says simply that we have no (considered) interest in interfering with people's liberties. Locke constrains the sovereign on "moral" grounds; Hobbes on prudential grounds. The question before us, is this: if there

exists claim-rights against the ruler interfering in the personal lives of the individuals, where do these claim-rights come from?

There are two options, as far as I can see. They either come from the agreements made by the people themselves or they exist prior to and independently of the agreements. If the latter, these Lockean claim-rights against liberty intervention exist as constraints on the types of covenants that can legitimately be made. But if the contractarian credo is as I have expressed in chapter one (namely: that what is rationally agreed upon by all people concerned is *necessarily* morally permissible), we see that a constraint on what counts as rational agreement goes *against* this contractarian tenet. It would mean that these claim-rights pre-exist the agreements. They exist in the State of Nature. For Hobbes, natural relationships are defined by might, not right, so the notion of "rights" in the State of Nature is inconceivable.

> To this warre of every man against every man, this also is consequent; that nothing can be Unjust. The notions of Right and Wrong, Justice and Injustice have there no place.[277]

---

[277] Hobbes, *Leviathan*, xiii (66). Elsewhere Hobbes appears to maintain that rights do exist in the State of Nature. "And because the condition of Man...is a condition of Warre of every one against every one...there is nothing he can make use of, that may not be a help unto him, in preserving his life against his enemyes; It followeth, that in such a condition, every man has a Right to every thing; even to one anothers body" (Hobbes, *Leviathan*, xiv, 67). Clearly this "right" does not entail a claim-right, by any means, and this quote too is sufficient to show that there is no room for Lockean restrictions on men's actions in the State of Nature.

If claim-rights pre-exist society, then the ontogeny of morality pre-exists society. Such a view is decidedly non-contractarian. For contractarians, morality is explicitly artificial, created by man and for man for the express purpose of social living.[278] By contrast, Locke's conception of the ontogeny of morality has a theological basis.

"A Hobbist" says Locke, "will not easily admit a great many plain duties of morality."[279] These "duties" that Hobbes fails to recognize hold only for those who believe in God.[280] The duty that Lockeans but not Hobbists are under, then, is derived simply from God's being our creator.[281] God owns us from having mixed his labour to create us. If I have a property right in x, that entails that I have a right to do anything I want with x. Having a right to do anything with x, entails a duty of others to refrain from preventing me from doing anything with x. This duty

---

[278] "For the Lawes of Nature (as Justice, Equity, Modesty, Mercy, and (in summe) *doing to others, as wee would be done to*,) of themselves, without the terrour of some Power, to cause them to be observed, and are contrary to our naturall Passions, that carry us to Partiality, Pride, Revenge, and the like" (Hobbes, *Leviathan*, xvii, 87).

[279] Locke, MS. quoted in John Dunn, *The Political Thought of John Locke*, Cambridge: Cambridge University Press, 1969, 218-219.

[280] Locke, "Letter Concerning Toleration," London, 1690. See also the discussion by Gauthier in *Moral Dealing*, 24.

[281] Locke, 2T, II, 6 (9).

of forbearance holds also for x should x be a person. And since this is the case with God having property rights in us, we have a duty to perform according to God's command.[282] For Hobbes, the explanation for obedience to God is simply fear.[283] We can clearly see, then, that Locke and Hobbes differ on their reasons for acquiescing to God's decrees: For Locke there is a moral reason; for Hobbes it is purely prudential. If we dismiss the notion of God, as Locke did not consider, then we lose the moral strictures against the invasion of property rights and individual liberties. The prudential grounds for the duty of toleration that Hobbes argues for remain unaffected.

For Locke, men have plain duties even in the State of Nature, much in the way that Butler saw it. There is a moral order already existing in the State of Nature and this includes the *duty* to refrain from preemptive war. For Hobbes, however, man must create a moral order out of the State of Nature. Even if we have shown, contra Hobbes's express beliefs, that it cannot be rational within the Hobbesian State of Nature to resort to preemptive violence, this is nevertheless an appeal to *rationality*, not duty.[284] Furthermore, should one have

---

[282] See also David Gauthier, *Moral Dealing: Contract, Ethics, and Reason*, Ithaca and London: Cornell University Press, 1990, 35.

[283] Hobbes, *Leviathan*, xi (54); xii (55).

[284] Hobbes, *Leviathan*, xiii (69). Some, however, claim that according to contractarianism, it is one's "duty" to act rationally (for example, A.E. Taylor, 35-36). Thus, duties exist in the State of

the ability to join forces to attack a singleton without threat of confederacy on the part of the defendant, Hobbesian psychology will decree it rational to do so, and no added strictures of duty, should there be any, can intercede. The task for Hobbes and contractarians in general is to construct a moral order to get us out of the State of Nature; for Locke, evidently, it is to discover an already existing moral order.

Far be it for me to deny that contractarians recognize claim-rights. The difference between Locke and Hobbes is how they ground these claim-rights. So far, I have attempted to show that contractarians in general, and Hobbes in particular, cannot ground claim-rights in the way that Locke does. What needs to be shown is whether contractarians can ground claim-rights at all, if they wish their theory to remain plausible. Claim-rights for contractarians are grounded in the nature of the agency-agreement itself. This, I believe, requires no further grounding than the fallback position. Claim rights are born from agreements. They do not exist prior to agreements. But if this is so, there is nothing in principle wrong with agreements to have one's liberty curtailed. There is no "ought" to it. Simply, it is *unlikely* that that is the sort of thing people *would* agree to. If so, Locke's and Hobbes's position amounts *in practice* to the same thing.

Nature for Hobbes as well. I wish to deny this interpretation in the following section.

Hobbes speaks abstractly: what is right is what people would rationally consent to, and this is left open-ended. Locke fills the content in: what people would rationally consent to is not to have their liberties curtailed (beyond some minimum, left unspecified for now).

The implications of these two views should not differ. If they do, there is some reason to mistrust contractarianism. Hampton, however, believes there are differences. Accordingly, Hobbes's fallback position has a "pragmatic" advantage over the Lockean one, while Locke's view rectifies the dangerous "dark" side of the Hobbesian stance.[285] Supposedly, the "dark" Hobbes view allows and even supports the abuse of minorities if that abuse is tolerated by the convention-producing majority who are unharmed by it or who may benefit from it.[286] But if claim-rights come about through the nature of the agreement, then the claim-rights against this sort of thing exist because the people themselves (particularly the convention-producing majority) have decided against it. But if they *decide* this, then it is not the case that such abuse would be tolerated by the majority. Another dark side for Hobbes's fallback position, Hampton presumes, is that the ruler can take as much advantage as possible over the ruled people.

---

[285] Hampton, 254

[286] Hampton, 254

This would in fact be "right" if they could get away with it, but morally wrong under Locke's ethics *even if they could get away with it*.[287] But if the rulers could get away with it, this implies the people are either not aware of it or do not care about it that much. Even if Locke is correct in pointing out a "moral" grievance, so long as Hobbesian psychology is correct, there would be little motivating force behind Locke's moral condemnation. If there were such force behind his moral outcry, that would only be so on the basis that the people have noticed the infraction and care about it -- but this is what is denied in Hampton's scenario. Thus, although Hobbesian or contractarian claim-rights come about through artifice and Lockean claim-rights through natural, religious, or other *a priori* means, the implications of the differing ontogeny should not differ in practice.

Many have recognized that Locke provides the moral content we want, but if we accept a contractarian basis of morality, we can not quite get these results. As David Gauthier, remarks, "modern moral philosophers want to begin with the [Hobbesian] conception of man as an individual appropriator and derive a Lockean, strongly overriding morality within the framework of Hobbist

[287] Hampton, 254.

secularism and autonomy,"[288] but that "such a morality is not to be found."[289] If Hobbes is correct, then it is true that there is nothing intrinsically wrong with a sovereign's curtailing the people's liberty in order to procure some desired good of the people whose liberty is curtailed. If it is not conceived to be a benefit, the ruler runs the risk of being ousted. The desired good achieved needs to be of more value to the individuals than the loss to their liberty. It may be the case that, in fact, any move that curtails one's liberty will not procure one's desired good. But this seems unlikely. Weaker, it may be the case that there will often enough be an alternative method of achieving that good without similar loss of liberty. Libertarians argue so. But even if it is so, this is a contingent matter; and it does nothing to criticize the *in principle* claim that violations of liberty may be morally permissible under a contractarian defence of ethics. And if this is true, then even though in practice contractarianism and libertarianism will appear similar, Lockean libertarianism can not be a *necessary* outcome of contractarianism. Libertarianism, after Locke, specifies what the *content* of a rational agreement will look like between the state and the people, and this will tend to be, according to libertarians, privatization. Contractarians, on the other

---

[288] Gauthier, 44.

[289] Gauthier, 27.

hand, whether agreeing with the libertarian applied ethics or not, are committed to leaving the *content* of any agreement unspecified. *Whatever* the agreement is, that is necessarily morally permissible. Should this be agreement for an absolute sovereign, or an agreement between the people and state to run an egalitarian regime, then this state is morally permissible; whether or not a different system would be better for them *in fact*. Contractarianism is open-ended. It is necessarily so due to the open-endedness of individual self-interest.

11.5 THE SCHISM

In this chapter we recognized a rift between Hobbesian and Lockean ethics concerning a restriction on the state's or sovereign's power. Locke believed there is an *a priori* moral reason for the state to be limited in power. Hobbes did not. It is useful to think of this distinction as that between *standard-bound* and *preference-bound* restrictions. The genesis of contractarian theory is to ground ethics on individual preferences. Contractarianism is committed to being a preference-based moral theory. As I have argued above, Locke veers from this strict contractarian grounding when he introduces standard-bound restrictions to the scope of the state's power. His motive for doing so is clear enough. Life under an absolute sovereign would be hell. It would still be hell even if it were slightly better on some scale than life in the State of Nature. Libertarians, in

particular, follow Locke on seeing the need for restricting the state's power in order to preserve individual (negative) liberty. Eradicating this situation by introducing a standard bound restriction on the power of the sovereign, though, is not permissible for a true contractarian. In order to serve the fundamental purpose of concerned parties, contractarianism must give rise to a choice-rule which is preference-dependent and standard-independent. Standard-bound contractarianism in the manner of Locke is inconsistent with the *raison d'etre* of contractarianism: to have social choices made under a rule which selects states of affairs for how well they are *liked* (preference-based rule) instead of for what they *are* (standard-bound rule).[290]

The reader who wishes to remain in the liberal tradition of narrow ethics and yet avoid subjugation to an absolute sovereign has two options: give up contractarianism in favour of Lockean libertarianism; or discover a way to remain true to preference-bound contractarianism and nevertheless avoid the fate of being devoured by lions as well as foxes. In the following section, I will argue for the latter option. In short, the distinction between Hobbes and Locke does not amount to showing that contractarianism is committed to an absolute sovereign the way Hobbes envisaged it. Contractarianism as a moral theory need not be,

---

[290] See chapters 7 and 8 in this thesis for the full argument in favour of preference-based contractarianism.

nor in fact should be, confined to the Hobbesian claim of absolute sovereignty of the state. And it can do so while remaining wholly preference-based.

11.6 PROPERTY RIGHTS

I have argued above that true contractarian principles forbid *a priori* restrictions on the content of agreements. If a content of an agreement is the erection of state controlled public goods, nothing in the contractarian principles can forbid *a priori* restrictions on individual liberty should this help implement public goods. So long as life under a system of property rights is preferable to a system without it, we have reason to want the domain of the state's power restricted. It seems inconceivable that we should allow social choice to rule what I can or cannot do with my property. Property rights ought to remain inviolable rather than being subject to public choice.[291] Property rights, thus, are a good candidate for domain restriction. Much of the appeal of libertarianism rests on this assumption. In fact, all of the liberal ethical theories hold this to a considerable degree.[292] Although the benefits of upholding property rights has been questioned by non-liberal theories, it is not one that I shall attempt to challenge

---

[291] De Jasay, for example, makes this point, 102.

[292] For a recent and clever defence of private property following Locke, see David Schmidtz, *The Limits of Government*, Boulder, Colorado: Westview Press, 1991.

here. Upholding the inviolability of property rights is consistent with non-domain-restricted, or preference-based, contractarianism.

So long as parties agree to leave property rights in the hands of public choice, preference-based contractarians are committed to allowing such property redistribution to occur. And where there is property redistribution, we may suspect that property rights of individuals are therefore not inviolable. The state's visible hands taking from one to give to another is apparently abrogating the property rights of the first. Disallowing state intervention in property rights is a case of domain restriction. This follows because upholding property rights would trump the sovereign's will. If contractarianism forbids domain restriction, it seemingly must acquiesce to the dissolution of property rights. And doing so severs contractarianism from other liberal ethical theories.

Fortunately for contractarians, this conclusion is inaccurate. It is true that if the sovereign has utter authority over all jurisdictions, then that must include the jurisdiction of property rights. Should people complain, Hobbes has merely to remind them that they themselves voted in the sovereign knowing full well that part of the sovereign's duties may well be the confiscation of otherwise private property. We must remember that the Hobbesian sovereign and all that the sovereign should do has been antecedently agreed upon. Thus, however much we grumble, it cannot be said that the sovereign seizes our property without our

consent. Since a key factor of contractarianism is voluntary consent of all parties concerned, to cite property rights as an example of domain-restriction misses the mark. In this sense, then, there is indirect consent to the sovereign's redistribution of property. If it is consensual, even if only in this indirect way, then it is not technically a violation of property rights, since the having of private property does not forbid consensual transfer of it.

Pointing out that the sovereign's running roughshod over your property is not "technically" a violation of your rights may not be satisfactory, I admit. Once we have established the Hobbesian or absolute sovereign, there is nothing he cannot do that we have not in this roundabout way "consented" to, and this seems to misuse an otherwise viable notion of consent.

Fortunately, contractarianism as it is modernly defended is not committed to Hobbes's depiction of an absolute sovereign. Inviolable private property may be defended on preference-based contractarian grounds without appeal to any *a priori* standard. So long as people typically care about property rights, this would be a thing they would prefer to sanctify; property would be what they would prefer to keep free of the state's hands. The fall-back Hobbes, noted earlier in this chapter, is a limited sovereign; he has not the authority over all aspects of individuals' lives, and that includes issues such as property rights. That the domain must be unrestricted to be true to the contractarian tradition does not

mean there must be an absolute sovereign that is unrestricted; simply, there is no *a priori* standard preventing the people to submit to a system where property rights are not inviolable. There is no natural right against it. The absolution (or prevention) of property rights is unlikely to be in the agent's interests. Hence, it is unlikely to be that to which agents will agree. This is partly an empirical claim. It also gains support from our understanding of why we have the need to erect moral institutions in the first place: to preserve individuals' interests in the face of others trying to preserve their own interests. According to Gauthier, "Systems of property and government are legitimized in terms of the consent they *would* receive from *rational* persons in a suitably characterized position of free choice."[293] This does not preclude someone's being rational who wants to give his property away. Part of the definition of property is the owner's right to give it away if he so desires.

The point is: private property can be defended either by standard-bound or preference-based ethics, and thus the desire for the preservation of property rights is not in any obvious way a vote for standard-bound ethics over preference-bound ones. Preference-based contractarianism holds that it happens to be in individuals' interests to preserve property rights and to forbid

---

[293] Gauthier, 53.

on those grounds, at least *prima facie*, a state's ability to supersede those rights. Standard-bound ethics defend property rights on grounds independently of those preferences.

11.7 SUMMARY

The point of this chapter is to accept Locke's modification of the contractarian position in regards to what Hampton has termed the "fallback position." I have attempted to argue that, nevertheless, we cannot get quite the results that Locke asserts from this fallback position. Further implications of this divergence will be taken up in the following chapter. I will investigate whether contractarians are committed to a centralized state, even if less than absolute sovereignty, or whether there is room for a decentralized minimal state within the contractarian tradition. Specifically, I will examine whether we can get a libertarian manifesto from occurrent contractarian grounds.

Chapter 12

WHY I AM NOT A LIBERTARIAN

Now that we have seen the moral theory of occurrent preference-bound contractarianism, it is well to wonder what its political implications are. Recall the discussion in chapter eleven. Contractarianism of the Hobbesian brand was committed to state intervention in all aspects of one's life. This was impalatable for Locke. To rectify it, Locke left preference-bound contractarian principles in order to justify a more limited state involvement. Can we arrive at Locke's more liberty-preserving conclusion through occurrent contractarianism? The answer is Yes. What matters for occurrent contractarians as much as for libertarians is that our individual liberties to pursue our individual goods is preserved as much as possible amenable to others' like liberties being preserved. Recognizing the contingency of one's liberty preservation on others' like liberties, however, creates a block to the recognition of full libertarian policies rather than inevitably leads to libertarianism. Or so I shall attempt to argue here.

12.1 LIBERTARIANISM AND CONTRACTARIANISM

Libertarianism is against centralized redistribution in general.[294] Contractarians, however, claim that what is moral is, roughly, what people agree to. Hence, for contractarians, there can be nothing wrong in principle with centralized distribution of goods, since it is in principle possible that the concerned parties will agree to it. It is for this reason that libertarianism cannot be founded on contractarianism, barring notable claims to the contrary.[295]

It is true, of course, that not all rational concerned parties agree to any particular redistribution program in today's North American society. Thus, one might assume occurrent contractarianism is not abrogated by applied libertarian policies.

For occurrent preference-based contractarians, morality serves a purpose, and that is to better enable individuals with disparate aspirations to attain their goals peacefully. Morality appeals to individual self-interest and the extent that it fails this is the extent to which it is no longer motivating. As individuals prosper within a moral realm as opposed to an amoral realm, it is in each person's self-interest to live in a moral domain. As the moral domain can not flourish without

---

[294] There are exceptions. For example, the Dominant Protection Agency of Nozick (Robert Nozick, *Anarchy, State, and Utopia*, New York: Basic Books, 1974, 15).

[295] Jan Narveson, *The Libertarian Idea*, Philadelphia, Temple University Press, 1988, 131; 154; 165-166; 175-184.

general obedience to the moral dictates of that system, it is indirectly in each individual's self-interest to obey the moral laws. And this follows even if the dictates curtail more directly self-serving goals. As the theory is decidedly self-interest based, there can be nothing intrinsically wrong with the possibility of curtailing one's liberty in order to procure some desired goal -- so long as it is desired more than one's liberty.

A libertarian may agree to this in principle, yet point out that, in fact, it will be in one's self-interest to abide by libertarian, free-market policies. If morality is founded on self-interest, then what provides more self-interested utility points, or "utiles," will be morally recommended. As libertarianism claims to offer the most individual utiles of any extant ethical theory, it contends to be the best ethical system. If so, libertarianism does follow contractarian grounds. Contractarianism presents the abstract formula; libertarianism fills in the content with empirical data. This, I think, is the standard interpretation of how libertarian principles are founded on contractarian ones. I hope here to demonstrate that it can not be so.

Part of the onus for libertarians is to show that the empirical support for libertarian policies is well founded. Whether this is so or not I confess ignorance. It requires specialization that I do not have. One requires the support of other sciences including economics, psychology, ergonomics, sociology, and perhaps history. Should these disciplines converge in pointing toward libertarian practices,

I think we should acquiesce. Whether they do or not, moreover, cannot be resolved by a few studies. We need also to be versed in statistics and general scientific method. Although philosophers like to pride themselves for being part of a discipline that founded most of the sciences, this does not qualify current philosophers as experts in these offshoot fields. This is beside my main thesis, however. My avowed ignorance in these applied matters does not affect the theoretical distinction I wish to make between contractarians and libertarians, for there is a theoretical problem with the above sketched libertarian response. *Contractarians need not adopt libertarian practices* even if *the experts unequivocally proved the teleological benefits of libertarianism.*

This may seem surprising. Have I contradicted myself? How can a theory claim to be interest-based, and yet refuse to adopt a system the consequence of which is beneficial on self-interested terms? The answer lies in what we have gleaned from our discussion of self-interest.

12.2 SELF-INTEREST REVISITED

Occurrent contractarianism is a preference-based ethic. The focus is on *occurrent* preferences rather than considered preferences for the reasons outlined in chapters four and five. Relying on considered preferences to ground ethics is relying on standards that are independent of the preferences individuals actually hold. To

claim these standards are the preferences individuals *ought* to hold is not to ground ethics in preferences. Our goal of making morality a rational pursuit, where rationality is individually motivating, determines, I have argued, a stringently preference-based theory of ethics. This stringent preference-based theory is occurrent contractarianism. Self-interest, on this model, is rightly to be defined by an individual's occurrent beliefs and not by anything else. Occurrent beliefs are the beliefs an individual happens to endorse. Furthering those beliefs or desires or preferences, whatever they are, will further the individual's self-interest, occurrently defined.

Hypothetical contractarianism, recall, or at least realist versions of hypothetical contractarianism, is concerned with what preferences individuals occurrently hold as well as with what preferences individuals would likely hold given specified circumstances and their occurrent beliefs and preferences. Libertarians claim that given people's occurrent preferences, they would most likely see the benefit of libertarian policies if given enough time to listen and understand the libertarian arguments. At first glance, this appears like an empirical claim. If so, we do not need philosophy to settle the issue. The claim need not be purely empirical, however. I believe libertarians conceive this claim to be more along the lines of a simple deduction: you want x (a variable to be filled in by the subject, and hence motivating to the requisite degree) and you can get more of x under this system,

since it permits you the utmost liberty to pursue your own goals and preferences. As put, what self-serving individual would not want to adopt libertarian policies? If it is true that *whatever* our goals are, these are more likely to be fulfilled under libertarian systems than any other, we would have to be irrational not to buy into libertarianism. It is more probable, surely, that there is a flaw in libertarian theory than that we are all irrational. One possible reason is because the variable x is often a different sort of commodity than material possessions. Security, for example, may be better purchased under a more egalitarian scheme than a libertarian one. At any rate, so long as people think that, and act on it, occurrent contractarians are not in a position to countermand that preference, however illguided it may be.

Security is a commodity that can be purchased along with anything else, libertarians are careful to remind us. Privatized insurance agencies will have a market if people are willing to spend money on security, and thus security, too, can be bought -- and quite likely at a cheaper rate than through systems of centralized redistribution. As I have already mentioned above, my goal here is not to decide whether this is empirically true. Our question is simply this: What if it is not convincing? What if people prefer that some decisions are made for them through a central organization. Even if it is more expensive in the long run, it is not inconsistent that people may nevertheless prefer to pay the price, not merely for the extra security, but also because of its *ease*. The hassle and the extra time in

searching out a better bargain may often not be worth the investment. In virtual reality, computers may have the expertise and time to survey and pick what always maximizes their programmed preferences. In the real world, however, we have often not the brains, nor the resources, nor the time, nor the energy, to maximize our every preference. So long as we can *satisfy* our preferences, however, we can be content. In fact, given our constraints, preference satisfaction, rather than preference maximization may well be the more rational method. My point here is that even if it is true that libertarians can tell us how to *maximize* our preferences, we may rationally decline given that our situation often lends itself to *satisfying* our preferences instead. Where people are willing to satisfy their preferences, by granting certain overarching functions to state legislation, say, we know that occurrent contractarians will of course not intercede. What are libertarians willing to do? If they agree with occurrent contractarians that voluntary contributions to permit centralized state control over certain social goods, there is then no fundamental difference between the two theories, and since libertarianism is supposed to be built on the foundation of contractarianism, it is a redundant theory, all its talk of negative rights notwithstanding. On the other hand, they may insist that the agreement is irrational in this case, and thus not to be considered binding. If so, libertarianism is opposed to contractarianism, and thus can hardly be claimed to be rooted in contractarianism. Occurrent contractarians are committed to not shoving

down anyone's throat what is not in their occurrently defined self-interest. Libertarianism, on the other hand, evidently must go beyond occurrent preferences in order to have people accommodate a libertarian schema -- even though what the libertarian schema professes is the liberty to pursue one's own occurrent preferences.

12.3 PARETIANISM

Libertarianism may be said to be founded on contractarian grounds with the assistance of adopting the Pareto principle. The Pareto principle claims that should any state improve at least one individual and not harm another it is to be preferred. Paretianism and egalitarianism are incompatible theses since paretianism permits inequality and egalitarianism does not. Imagine two states, S_1 and S_2, comprised of ten individuals each. In S_1, everyone has $10. In S_2, everyone has different amounts, although the poorest member (still) has $10. Moreover, in S_2 there is an increment of $10 between each citizen. The richest has $100. Due to the inequality in S_2, egalitarians would prefer S_1 (with a mean of $10) to S_2 (with a mean of $55). Followers of Pareto, on the other hand, would prefer S_2 over S_1 since a mean of $55 is clearly superior to a mean of $10 and no one is worse off strictly speaking in S_2 compared to S_1 (i.e., S_2 is pareto superior over S_1). As it is the case that S_2 is clearly more beneficial to 90% of the population over S_1, and S_2 is also endorsed by

paretianism, it would appear that the logic of paretianism is preference-based. Partly for this reason, paretianism is endorsed wholeheartedly by libertarians.

Where changes are being considered between different social and political structures, paretianism runs counter to occurrent contractarianism. I do not mean that contractarians would necessarily prefer S_1 to S_2 in the above scenario. One can dispense with paretianism without being an egalitarian, and this is what I intend to do.

The emphasis on human psychology in ethics is an important facet of much of recent ethical thought. Owen Flanagan has termed this reliance on psychology, "psychological realism."[296] Contractarianism is much indebted to this notion of psychological realism. Hobbes was very specific about the sorts of people we are, and was determined to build an ethical system given the groundwork of human psychology. The very goal of making ethics appeal to self-interested people is committed to being as accurate as possible about what sorts of people we are. If we abide by psychological realism in other cases, we must abide by it in all cases if we wish to be consistent. Among the general characteristics humans have been discovered to possess, one in particular is going to make paretianism difficult to sustain; and that is "relative deprivation." Relative deprivation is a psychological

---

[296] Owen Flanagan, *The Varieties of Moral Personality*, Cambridge Mass: Harvard University Press, 1991, 32-55.

state that people sometimes suffer for being made to appear less well off by the greater advantage of another.[297] Paretianism claims that so long as at least *one* person is better off in S_2 than S_1 and no one is worse off, S_2 should be adopted over S_1. In S_1, recall, everyone has exactly $10. Imagine a new state S_2^*.[298] In S_2^*, we shall imagine that only one person has more than $10. Let's say, he has $50, and everyone else still has $10. If relative deprivation is a psychological fact for a significant number of people, it cannot be the case that S_2^* will be preferred to S_1 by anyone other than the benefiting party. And should we adopt a theory that makes it our *duty* to prefer S_2^* to S_1, we have left a preference-based ethic in favour of a standard-based ethic. Libertarians, in endorsing paretianism at the level of social and political structural changes, override preferences, at least those preferences affected by relative deprivation.

---

[297] For early studies, see J.A. Davis, "A Formal Interpretation of the Theory of Relative Deprivation," *Sociometry*, 1959, 22, 280-296; and W.G. Runsiman, *Relative Deprivation and Social Justice: A Study of Attitudes to Social Inequality in Twentieth Century England*, Berkeley: University of California Press, 1966. During the 1960s two of the most serious riots in history by American blacks did not take place in the areas of greatest poverty, but in Watts and Detroit, where, compared to the whites, things were bad for the blacks. For a more recent review, see M. Bernstein and F. Crosby, "An Empirical Examination of Relative Deprivation Theory," *Journal of Experimental Social Psychology*, 1980, 16, 442-456.

[298] I prefer the convention of using S_2^* over a third state S_3, because S_3 may give the unwanted implication that there is yet another choice to be made, i.e., S_2.

Let me be clear, for in one sense what I have said above is false. If S_1 is the present society, and S_2^* is exactly the identical present society with the one exception that Jones is $40 richer through voluntary transaction with Smith, the other members of society must indeed accept the pareto improvement. They have no right to complain. That they feel they should have the money rather than Jones cannot count as sufficient justification, as put, to prevent Jones's prosperity. Their complaints, if any, are empty. That they must acquiesce to Jones's riches does not mean they need to do anything. It is a negative demand. They must not interfere with the voluntary transaction between Smith and Jones. At the level of free market transaction, paretianism is wholly endorsed by occurrent contractarianism. In fact, paretianism at this level is a condition of contractarianism. Recall in chapter one that although the existence of negative externalities suffices to halt or alter agreements, meddling by busy bodies outside the circle of relevance cannot affect the legitimacy of voluntary, uncoerced agreements between concerned parties.

Where paretianism is peculiar to an occurrent preference model is at the level of societal and political structural changes. If S_1 is the present political structure, and S_2 is a pareto superior but completely different political structure, then I resist the claim that people *must* accept S_2. Even if S_2 is pareto superior, I may not wish to undergo the necessary changes to help bring it about. And this is still the case even when the costs to me to change from S_1 to S_2 are outweighed by

the benefits assumed under the conditions of pareto superiority. In other words, if the shift from S_1 to S_2 demands positive action on my part (for example abiding by the different terms of the new political structure) and especially where I receive no net benefit (albeit I suffer no loss either given the assumption of pareto superiority), then occurrent contractarians cannot accept the mandate of paretianism. On the other hand, if no action is required by me save my non-interference, pareto improvements indeed must be endorsed. Changes within the free market do not demand positive actions from those outside the agreements. Changes of political structures do.

Some may contest that it is not so much one's duty to adopt any pareto superior state; simply it is rational to do so. If a change from S_1 to S_2 is indeed pareto superior, then surely rationality cannot object to the change. If this is so, then relative deprivation is an *irrational* psychological state and should not be counted as one's *proper* preference. But then preferences and rationality part ways and we need to decide whether ethics should be sensitive to rationality, strictly defined, or preferences. I have already argued for why I believe we should adopt a preference-based ethic, rather than a standard-bound one, as would be the case for a rational ethical model that ignores certain occurrent preferences. We do not run this danger here. In point of fact, what is irrational is preferring S_2^* over S_1, not the reverse. If I am neither better nor worse off in S_1 or S_2^*, I, of course, should be

indifferent. That is to say, I should be indifferent if we assume non-tuism in our quest to ground ethics on non-moral assumptions. Accepting the premise that people are non-tuistic is not meant as a psychological claim. Its purpose is to rule out the assumption that people will cooperate when it is less than rational for them to do so simply because they have some affection for others' well-being. Natural affections for others is what cannot be counted on. By ruling out the influence of others' bargaining by affective feelings, one is forced to interact with others on a strictly rational level. People are motivated to prosper for themselves. When changes do not involve improvement over an individual's status quo (as is the case for nine tenths of the subjects in our scenario), there is *no* motivation to choose that situation. We are dealing with individuals motivated by their own individualistic occurrent preferences. If rationality cannot decide between protecting my finger from a paper cut to the extermination of the planet, there is also going to be nothing inherently rational in my preferring S_2^* to S_1.

One might argue that choosing the pareto superior state may still be a rational act even if I, personally, am indifferent between the two states. Since I am indifferent, I can rationally delegate my choice selection to the person who cares, i.e., the lone individual who actually prospers under S_2^*. For ease, let us imagine it is you. As you prefer S_1 to S_2^*, S_1 gets chosen, and this should be fine for me. In this case, I prefer what you prefer. That I may rationally do this, however, does not

mean that the *uniquely* rational thing for me to do is to delegate my preference to you. And so long as I am affected by relative deprivation, it may not even be a rational choice. If I am indifferent between S_1 and S_2^*, this should mean I lose nothing in the change from S_1 and S_2^*. But it is not necessarily true that I lose nothing. Perhaps you're preferring S_2^* over S_1 will actually affect my indifference. If I am concerned with relative standing, for example, then you're preferring S_2^* over S_1 may itself tip the scales toward my preferring S_1.

Egalitarian doctrines jump on the relative deprivation bandwagon to claim that *therefore*, we should not allow others to have more than anyone else, at least within an undefined range.[299] I believe this goes too far. Imagined suffering from relative deprivation is a petty emotion and should not be sufficient to halt beneficial social change any more than pent up violence should be allowed free reign simply because it is a psychological fact. I think this response is absolutely correct, but we should be wary of its limitations. What the response does not do is support paretianism. What it shows is a reason to allow people a liberal degree of free reign. If Barney beside me gets $10,000 more than me, I may feel disappointed in comparison. But my disappointment is irrelevant. Barney's fortune has nothing to

---

[299] See, for example, Robert Frank's discussion of the social implications of relative deprivation in *Choosing the Right Pond: Human Behaviour and the Quest for Status*, New York: Oxford University Press, 1985, eg. 5-7.

do with me (we assume for argument's sake[300]). Barney does not need my vote on whether he gets an extra $10,000 (unless he gets it from me, which is unlikely). Asking for my vote is precisely what paretianism asks. Barney may contest: "What the hell does he have to do with this? He, being a psychological being, will probably resent it. But his resentment is out of line. He is not one of the concerned parties in the contracts that I have made to earn this extra $10,000, and since contractarianism requires only the votes of those party to the agreement, whether he feels relatively deprived as a result of my good fortune is inconsequential." Barney's response is absolutely right. It shows that contractarianism permits inequalities. It also shows that paretianism is an unnecessary appendage to, if not downright antithetical to, contractarianism. Let me explain.

People will prefer that others are not better off than themselves, or at least there are enough psychological studies pointing to this rather dour fact, and so the paretian claim that people should prefer S_2^* to S_1 is psychologically false. It is irrelevant, however, whether they prefer S_2^* to S_1 or S_1 to S_2^*. All that matters is whether they are one of the concerned contracting parties. If not, their feelings of deprivation are inconsequential to social theory.

---

[300] Of course it may be the case that Joe's prosperity has a lot to do with me. Perhaps he robbed me. Then, my concern does matter. In the discussion above, this is not the situation we are considering.

If they are part of the contract, their feelings of relative deprivation may get in the way of real benefits, and the individuals themselves have to accommodate that in their appraisal of any contracts. Robert Frank argues that the importance of relative standing is itself a commodity -- bought and sold as with everything else.[301] Thus, we may assume, a pareto superior position will be chosen by those who do not benefit by the change only so long as they are *compensated* by the one(s) who do benefit. Frank is partly right about this, but it is obscured by the fact that he makes it sound as though what is required is a social compensation program run through a centralized distribution office. Simply, if Jones wants to make a deal with Smith in order to reap an extra $10,000, Smith is not likely to be merely indifferent about it if Jones's proposal leaves Smith at the status quo. Smith will likely want something in return. Jones cannot be blind to that in his proposal. If Jones requires some action from Smith in order to reap his fortune, he will need to make it worth Smith's while to help him. Jones's telling Smith about paretianism, alone, is unlikely to do the trick.

Just as Smith will unlikely be convinced by paretianism, so too contractarians should have no use for Pareto, at least concerning claims toward structural changes of society.

---

[301] Frank, 108-109, eg.

12.4 OCCURRENT-BASED LIBERTARIANISM

Standard-based contractarianism appears to be inconsistent with what we know as libertarian policies. For example, libertarianism wishes to claim that suicide is permissible so long as no harm to others is committed, and the act is uncoerced. This is in line with *occurrent* contractarianism, but not standard-based or considered contractarianism. Basing libertarianism on *occurrent* preferences, however, meets with the following difficulties. In our present society, people generally *agree* with taxation. People in general do *not* wish to abandon social security and health care and building regulations and public education and public road systems. That libertarianism nevertheless advocates the *benefits* of abandoning centralized programming can not be a *moral* matter, then, so long as the libertarian foundation is occurrent contractarianism. It is, instead, a matter for economists and political scientists. It would not be even a philosophical issue as far as I can tell. On *occurrent* contractarian grounds, whatever is agreed to *is* morally permissible. Thus, so long as libertarianism is based on *occurrent* preferences, libertarians *cannot* recommend our abandoning certain agreed upon policies on the grounds of its being *immoral*. That is to say, they can recommend libertarian policies, but not on moral grounds. Claiming that taxation is theft on these lines would be merely a rhetorical device. Theft is immoral but taxation is not so long as the individual members of a society do happen to agree in principle with it.

However many do agree in principle with taxation, there may well be a number who do not. Taxing these people countermands their occurrent preferences. The libertarian claim that tax is theft is true for these people and it would appear that occurrent contractarians are committed to agree. A plan to exclude these conscientious tax evaders from the costs as well as from the benefits (if any) of taxation introduces free-rider problems. If the roads are paid through tax money, how can we detect and prevent non-tax-payers from using the roads? The libertarian answer is to simply do away with taxation on the whole and privatize roads and education and health-care and what-not. Under privatization, only those who benefit pay, and only those who pay benefit.

Quite often it is only poor imagination that prevents people from seeing how cost-efficient privatized goods can be offered. Nevertheless, the libertarian scheme has been built around the minority. What if, we ask, the majority prefer a centralized distribution system of goods such as education, health care, roads, and the like. Arguing which system is most cost-effective is beside the point for an occurrent-based ethic. Rather, we must see which system, if any, violates the occurrent contractarian model of morality. Libertarians argue that the centralized distribution system violates the preferences of the minority who wish not to be taxed. Does the libertarian schema violate the occurrent interests of those who agree in principle with centralized distribution of goods? Moving from a taxation system to a

libertarian system, even if the move is pareto superior, cannot be demanded of individuals unless harms are committed in the present system that can be avoided under the new system. So long as there are some who prefer life under a centralized distribution structure than one under a libertarian privatized structure, then occurrent contractarianism cannot advocate the move.[302]

What we may well disagree with is the use our tax money is put to. I may agree to a number of things, but not necessarily to all uses. Since taxation is forced and yet some of that money is distributed in ways counter to my interests, at least part of taxation violates the contractarian principles of morality and justice. Granted, but what of the rest of the taxation money? That we can disagree with how some of our tax money is being spent is not to disagree with taxation in general. And the point where local dissatisfaction degenerates to wholesale dissatisfaction with taxation and centralized distribution is where libertarianism leaves occurrent grounds.

One could allude to the fact that we generally have no choice in paying our taxes, or how that tax money is to be spent, to disabuse ourselves form the view

---

[302] Of interest is to note that occurrent contractarianism favours conservatism. If the present structure was a libertarian system, occurrent contractarians would be reluctant to move to a centralized state system. Occurrent contractarianism places the onus on the reformers to show that either there is no real agreement or that negative externalities exist under the present system before occurrent contractarians can advocate changing from one political structure to another.

that forceful taxation meets the requirements of mutual consent under occurrent contractarianism. This ignores our having established the legitimacy of commitment under an occurrent preference model.[303] It is not a violation of Ulysses's occurrent preferences to refuse to untie him from the mast as he pleads for release sailing past the sirens because he made public his overriding occurrent preference not to be unbound. Most goals have costs. We are not likely to prefer to pay these costs all by themselves, but some of them we tolerate in order to satisfy our goal. Morality as an institution is like this. We give up some of our predatory actions in order not to be the prey of others. Political structures are no exception to this rule. We willingly make some sacrifices for what we assume to be a larger benefit. That we gripe during the time of the sacrifice, while we pay our taxes, is not to say we therefore have not consented to this, any more than ignoring Ulysses's gripes indicates his crew is violating his occurrent preferences.

There is a serious worry, of course. To what extent have we agreed to suffer for the benefits of a political structure? Certainly we have rights to complain we are paying too much, or that we are receiving the wrong sort of compensation, or are receiving too little benefit for the cost. Libertarians are wise to alert us to these dangers. One need not be a libertarian to make these remarks, however. We can

[303] See chapter 8 in this thesis.

criticize the internal workings of the government without abandoning the whole structure.

12.5 STANDARD-BASED LIBERTARIANISM

What the above reveals is that libertarianism at its full theory can *not* be based on occurrent contractarianism and be consistent, where its "full theory" advocates a fully privatized state. It is in our occurrent interest to disallow a Hobbesian absolute sovereign, but there is much room on the continuum between that Leviathan and a Nozickean minimal state. There is a possibility that libertarianism is based on considered contractarianism, however. Under a considered contractarian rubric, there need not be such a reliance on what people actually have or would consent to given their occurrent preference structures. Rather, what is in the people's interests is determined on the basis of external principles. Moreover, there is something quite plausible in this supposition. If moral permissibility is not based on what preferences and agreements people *happen* to have or make, but on those preferences and agreements people *should have* made, if they were right thinking and suitably placed, this will allow the claim that libertarianism is the necessary outcome of a *moral* theory. That people *in fact* individually prosper under a libertarian system, whether they know this or not, will be grounds to institute a

libertarian order, since what is morally permissible is what furthers individual interests at least hypothetically defined.

As argued in chapter four, standard based contractarianism is the *wrong* sort of contractarianism. In brief, it ignores preferences in favour of antecedently held (and in this case highly contentious) standards of excellence. There is another problem, however, with basing libertarianism on standard-bound contractarianism that is independent of its being merely the wrong sort of contractarianism. Libertarians are inconsistent on their grounding of contractarian principles. Although they require considered or standard-based contractarianism in order to wish to adopt libertarianism when the occurrent preferences of the people see the benefits of non-libertarian doctrines, libertarians must also appeal to occurrent contractarianism. Standard-bound and occurrent contractarianism are incompatible, however.

It is important to understand that contractarianism restricted by considered preferences (we'll call this "considered contractarianism" for short) and occurrent contractarianism cannot be held at the same time. The implications of the two views differ. Take, for example, mass suicide. So long as all members voluntarily consent to kill themselves (i.e., there was no hypnosis, brain-washing, coercion, blackmail, and the like), occurrent contractarians can find nothing immoral about it. Given that this is their uncoerced preference, and their actions are more important to them

than the adverse effects these actions have on others, it would be morally permissible. And this is so even when it does not seem to be in their best interest from a god's eye-view. We may believe that almost any other decision would be better for them than suicide. I am not thinking of mercy killing. Perhaps the group believes the only way to save the world is through mass suicide on March 11, at 4:00 p.m. Or perhaps they believe mass suicide will give them immortality. We are free to scoff at their belief system. We think it would be more in their self-interest to be disabused of such crack-pot ideas. The point is, whatever their occurrent preferences, mass suicide is not their *true* self-interest, and for this reason considered contractarians have grounds to interfere. For we imagine that if they stood back and saw the state of affairs the way we see it, i.e., had they "considered" their preferences with a more objective perspective, they would see the ludicrousness of their plan. There are no such grounds for occurrent contractarians. Occurrent contractarianism, recall, makes no distinction between "considered" and "unconsidered" preferences, so long as these are identifiably the preferences the actors actually hold. To speculate on what would be their "considered" preferences is akin to the unrealistic or preference-insensitive versions of hypothetical contractarianism. We would be asking them to place themselves in a situation that is counter to their present state.

Hypothetical contractarianism legitimately adheres to occurrent preferences when we are considering the likely reactions of others to our behaviour. We must speculate on their likely preferences, for their occurrent preferences presumably have not taken into account our actions as of yet, and so do not count *for that reason*. Speculating on whether someone is acting under considered preferences or not is a different matter. Here, we know what their occurrent preferences are. In our example, it is to commit suicide en masse. This is their preference, and we are wondering whether we should interfere or not. To intercede on the grounds that if they were more like us, they would want to be stopped is to ignore their occurrent preferences in favour of the preferences we imagine we would have if we were in their same situation. Not only does this foist our beliefs on them, but it is also logically misguided. For *if* we were in their same situation, then it would be more likely that our occurrent preferences would be the same as theirs, and not as ours are presently in a counter-situation.

The purpose of this example is to illustrate that moral answers will differ depending on whether one is an occurrent contractarian or a considered contractarian. We need not agree on which is better to recognize that to adopt both is contradictory. Mass suicide would be at the same time both moral (on occurrent grounds) and immoral (on considered grounds), and any theory that tries to say both is therefore no theory at all.

If it can be shown that libertarianism relies indiscriminately on both occurrent and considered versions of contractarianism, we can deny the link between contractarianism and libertarianism. My short claim here is that libertarianism *does* rely on both versions.

Although libertarianism must be grounded on a standard-based contract, at least as far as encouraging others to alter their occurrent preferences to go along with libertarian policies, they explicitly admit that only occurrent preferences matter. The following sort of libertarian remark is clearly an *occurrent* contractarian credo: "So what if Scrooge doesn't want to give to charity; that's his choice." There is certainly nothing here that asks, "yes, but what is *really* in Scrooge's interests?" Butler, who also advocates that people should do whatever is in their *true* self-interest, is not typically thought to be a chancellor of libertarianism.

Let me summarize my main points up to now. Libertarianism can not be based *solely* on occurrent contractarianism. For, if so, the fact that people *do* agree in principle with taxation would show that there can be no claim that taxation is immoral. There can be no *moral* grounds for moving from our present political structure to a libertarian one. But neither can libertarianism be based *solely* on standard-based contractarianism. Standard-based contractarians will agree that the morally permissible political structure is what right-minded people, suitably placed, would consent to. And it may well be a matter for academicians to decide what that

is. Should it turn out to be libertarianism, then there is, seemingly, a grounding for libertarianism on standard-based contractarianism. This line of argument fails, however, for it cannot account for allowing Scrooge to do whatever he likes even if it is not in his best interest, coolly considered. It can not account for the permissibility of suicide. It can not account for the reliance on individual liberty to let people decide their own interests and goals, however stupid they may appear to others. It abandons, in fact, the very notion upon which contractarianism is based and to which libertarianism explicitly appeals.

Given these considerations, it appears that libertarianism requires both standard-based and occurrent forms of contractarianism, and that libertarians indiscriminately vacillate between the two. This is not a good foundation if the two interpretations of contractarianism are incompatible, as I have argued.

A libertarian objection at this point may be to insist that standard-based and occurrent contractarianism are not, at least necessarily, incompatible in the way that I claim. It is not inconceivable that what would be in the best interest of the people is to let people decide for themselves. That they decide for themselves shows the occurrent contractarian basis, and that this is the best political scenario shows it is wholly compatible with a standard-based contractarian basis. Simply put, what people would rationally agree to if suitably placed is a system that would allow people to pursue their occurrent interests. As nothing is inconsistent in imagining

this, libertarians do very well, thank you, in straddling the two contractarian interpretations. Thus, my complaint that libertarians are inconsistent in their theoretical support is mistaken. Or so this line of objection would assert.

I do not believe so. This neat mix of standard-based and occurrent contractarianism, as put, amounts to simply showing the merits of occurrent contractarianism. The claim here is that by allowing Scrooge to do whatever he pleases will likely yield consent from Scrooge. His assent is not based on any standard concerning what he would consent to if suitably placed. It is, rather, a tautology. People will consent to do what they want to do since, by definition, they want to do it. There is no reliance on the standard-based model. There is no provision that we will disallow Scrooge's miserliness on the grounds that if he were right thinking he would be more generous. Likewise, there is no provision that forbids Mother Theresa's excessive altruism. Nor is there any provision for criticizing the rest of us who acquiesce to taxation, however begrudgingly. That we would agree to live in a system that would honour our agreements is simply to point out the viability of occurrent contractarianism. Standard-based contractarianism is inconsistent with this. To ask whether people would agree with a system that would decide for them what counts as being in their interest is not likely to get actual consent. We cannot move to the legitimacy of standard-based contractarianism without the occurrent consent to do so. Interestingly enough, this is precisely the

form of argument that libertarians have raised against Rawls.[304] However plausible it is to choose egalitarian systems once we are behind the veil of ignorance, Rawls forgets that the bulk of us have little occurrent incentive to willingly step behind this veil.

12.6 SUMMARY

Good consequences are defined by the individuals themselves on contractarian grounds. The consequentialist aspect of contractarianism, then, does not mean that contractarians will necessarily adopt whatever political system leads to the best, objectively defined, results. Given this, it is beside the point for libertarians to tell us what those best results are -- at least, if their intent is to show that libertarianism is grounded on contractarianism. It is another matter altogether whether they want to educate us to seeing the light. This is certainly permissible, if not commendable. My point here, and throughout, is that there is simply no clear link between the two theories, save some overlap in policy endorsement.

What counts as good consequences is not determined by some philosopher or economist in his hermetically sealed tower. It is decided by the individuals themselves. Besides the large claim that libertarianism *is* a better system (as if that

[304] See Narveson, 132; 154.

were not enough), for contractarians to adopt libertarian policies, it must also be shown that the people involved believe it to be to their advantage. The people involved, and that means society at large, must *agree*. Otherwise, it is not conceived to be in their benefit -- whether or not it is *in fact*. And *this* is the problem libertarians face when they make bold the claim that their policies are grounded on contractarianism.

BIBLIOGRAPHY

Ayer, A. J., *Language, Truth, and Logic*, New York: Dover Publications [1946] 1952.

Baglan, Thomas, "Effects of Interpersonal Attraction and Type of Behaviour on Attributions," *Psychological Reports*, 48, 1981, 299-304.

Baier, Annette, "Doing Without Moral Theory?" in S. Clarke and E. Simpson (eds.) *Anti-Theory in Ethics and Moral Conservatism*, Albany, NY: State University of New York Press, 1989, 26-48.

Baier, Kurt, "Rationality, Value, and Preference," *Social Philosophy & Policy*, 5, 2, 1988, 17-45.

Baier, Kurt, *The Moral Point of View: A Rational Basis of Ethics*, Ithaca: Cornell University Press, 1958.

Baumrin, Bernard (ed.), *Hobbes's Leviathan: Interpretation and Criticism*, Belmont, Cal: Wadsworth Publishing Co., 1969.

Bernstein, M. and F. Crosby, "An Empirical Examination of Relative Deprivation Theory," *Journal of Experimental Social Psychology*, 1980, 16, 442-456.

Blackburn, Simon, "Moral Realism," in J. Casey (ed.) *Morality and Moral Reasoning*, London: Methuen, 1971.

Bradley, Francis Herbert, "My Station and its Duties" in *Ethical Studies*, London, 1876.

Brandt, R.B., *Ethical Theory*, Englewood Cliffs, NJ: 1959.

Broome, John, *Weighing Goods: Equality, Uncertainty and Time*, Oxford: Basil Blackwell, 1991.

Brown Jr., Stuart, "Hobbes and the Taylor Thesis," in Bernard Baumrin (ed.), *Hobbes's Leviathan: Interpretation and Criticism*, Belmont, Cal: Wadsworth Publishing Co., 1969.

Butler, Joseph, "Upon Human Nature" in *Fifteen Sermons at the Rolls Chapel* [London, 1726] reprinted in Bernard Baumrin (ed.) *Hobbes's Leviathan: Interpretation and Criticism*, Belmont, CA: Wadsworth Publishing Co., 1969, 16-25.

Campbell, Richmond, "Background for the Uninitiated," in R. Campbell and L. Sowden (eds.) *Paradoxes of Rationality and Cooperation: Prisoner's Dilemma and Newcomb's Problem*, Vancouver: University of British Colombia Press, 1985, 3-41.

Campbell, Richmond, "Sociobiology and the Possibility of Ethical Naturalism," in David Copp and David Zimmerman (eds.), *Morality, Reason and Truth*, Totowa, NJ: Rowman & Allanheld, 1985, 270-296.

Clarke, Stanley, and Evan Simpson, *Anti-Theory in Ethics and Moral Conservatism*, New York: State University of New York Press, 1989.

Danielson, Peter, "Evolutionary Models of Cooperative Mechanism: Artificial Morality and Genetic Programming." To appear in P. Danielson (ed.) *Modelling Rationality, Morality, and Evolution*, New York: Oxford University Press, 1995.

Danielson, Peter, "Evolving Artificial Moralities: Genetic Strategies, Spontaneous Orders, and Moral Catastrophe." Prepared for the conference on "Chaos and society" at the Université du Québec à Hull, June 1-2, 1994.

Danielson, Peter, *Artificial Morality: Virtuous Robots for Virtual Games*, London: Routledge, 1992.

Danielson, Peter, "Closing the Compliance Dilemma: How it's Rational to be Moral in a Lamarkian World" in Peter Vallentyne (ed.), *Contractarianism and Rational Choice: Essays on David Gauthier's "Morals by Agreement,*" New York: Cambridge University Press, 1991, ch. 16, 291-322.

Darley, J.M. and C.D. Batson, "`From Jerusalem to Jericho': A Study of Situational and dispositional Variables in Helping Behaviour," *Journal of Personality and Social Psychology*, 27, 1973, 100-108.

Darley, J.M. and B. Latané, "Bystander Intervention in Emergencies: Diffusion of Responsibility, *Journal of Personality and Social Psychology*, 8, 1968, 377-383.

Davis, J.A., "A Formal Interpretation of the Theory of Relative Deprivation," *Sociometry*, 1959, 22, 280-296.

Davis, Lawrence, H., "Is the Symmetry Argument Valid?" in R. Campbell (ed.) *Paradoxes of Rationality and cooperation*, 251-263.

de Jasay, Anthony, *Social Contract, Free Ride*, Oxford: Clarendon Press, 1989.

Dewey, John, *Theory of the Moral Life*, New York: Holt, Rinehart and Winston [1908] 1960.

Dion, Karen, "Physical Attractiveness and Evaluations of Children's Transgression," *Journal of Personality and Social Psychology*, 24, 1972, 207-213.

Dworkin, Gerald, *The Theory and Practice of Autonomy*, Cambridge: Cambridge University Press, 1989.

Ewin, R. E., *Virtues and Rights: The Moral Philosophy of Thomas Hobbes*, Boulder: Westview Press, 1991

Fishkin, James, "Bargaining, Justice, and Justification: Towards Reconstruction," *Social Philosophy & Policy*, 5, 2, 1988, 54.

Flanagan, Owen, *The Varieties of Moral Personality*, Cambridge Mass: Harvard University Press, 1991.

Frank, Robert, *Choosing the Right Pond: Human Behaviour and the Quest for Status*, New York: Oxford University Press, 1985.

Frankfurt, Harry, "Freedom of the Will and the Concept of a Person," *Journal of Philosophy*, 68, 1971, 5-20.

Frankfurt, Harry, "The Importance of What We Care About," in H. Frankfurt, *The Importance of What We Care About*, Cambridge: Cambridge University Press, 1988.

Frankfurt, Harry, "Identification and Wholeheartedness," in H. Frankfurt, *The Importance of What We Care About*, Cambridge: Cambridge University Press, 1988.

Gauthier, David, "Assure and Threaten," *Ethics*, 104, 1994, 690-721.

Gauthier, David, *Moral Dealing: Contract, Ethics, and Reason*, Ithaca and London: Cornell University Press, 1990.

Gauthier, David, "Morality, Rational Choice, Semantic Representation: A Reply to My Critics," *Social Philosophy and Policy*, 5, 2, 1988.

Gauthier, David, "Moral Artifice," *Canadian Journal of Philosophy*, 18, 2, 1988, 381-418.

Gauthier, David, *Morals by Agreement*, Oxford: Clarendon Press, 1986.

Gauthier, David, "David Hume: Contractarian," *Philosophical Review*, 88, 1979, 3-38.

Gass, William, "The Case of the Obliging Stranger," *The Philosophical Review*, 66, 1957, 193-204.

Gert, Bernard, "Hobbes and Psychological Egoism," in Bernard Baumrin (ed.), *Hobbes's Leviathan: Interpretation and Criticism*, Belmont, Cal: Wadsworth Publishing Co., 1969.

Glasser, William, *Reality Therapy: A New Approach to Psychiatry*, New York: Harper and Row, 1965.

Hampton, Jean, "The Failure of Expected-Utility Theory as a Theory of Reason," *Economics and Philosophy*, 10, 1994, 195-242.

Hampton, Jean, "Two Faces of Contractarian Thought," in Peter Vallentyne (ed.), *Contractarianism and Rational Choice: Essays on David Gauthier's "Morals by Agreement,*" New York: Cambridge University Press, 1991, ch. 3, 31-55.

Hampton, Jean, "Equalizing Concessions in the Pursuit of Justice: A Discussion of Gauthier's Bargaining Solution," in Peter Vallentyne (ed.), *Contractarianism and Rational Choice: Essays on David Gauthier's "Morals by Agreement,*" New York: Cambridge University Press, 1991, ch. 10, 149-161. Excerpted with minor modifications from "Can We Agree on Morals?" by Jean Hampton, *Canadian Journal of Philosophy*, 18, 1988, 331-356.

Hampton, Jean, *Hobbes and the Social Contract Tradition*, Cambridge: Cambridge University Press, 1986

Hansson, Sven Ove, "Money-Pumps, Self-Torturers and the Demons of Real Life," *Australasian Journal of Philosophy*, 71, 1993, 476-485.

Hardin, Russell, "Bargaining for Justice," *Social Philosophy & Policy*, 5, 2, 1988, 65-74.

Hare, R.M., *Freedom and Reason*, Oxford: Oxford University Press, 1963.

Hare, R.M., "Ethical Theory and Utilitarianism," In *Contemporary British Philosophy*, H.D. Lewis (ed.), London: Miller and Unwin, 1976. Reprinted in J.P. Sterba (ed.), *Contemporary Ethics: Selected Readings*, Englewood Cliffs, NJ: Prentice-Hall, 1989, 252-264.

Harman, Gilbert, *The Nature of Morality*, Oxford: Oxford University Press, 1977.

Harman, Gilbert, "Rationality in Agreement: A Commentary on Gauthier's `Morals by Agreement'," *Social Philosophy & Policy: Gauthier's New Social Contract*, 5, 2, 1988, 1-16.

Haworth, Larry, *Autonomy: An Essay in Philosophical Psychology and Ethics*, New Haven: Yale University Press, 1986.

Hobbes, Thomas, *De Cive*, Sterling P. Lamprecht (ed.) New York [1642] 1949.

Hobbes, Thomas, *Leviathan*, Buffalo, New York: Prometheus Books [1651] 1988.

Hume, David, *A Treatise of Human Nature*, L.A. Selby-Bigge (ed.) 2nd edition revised by P.H. Nidditch, Oxford: Oxford University Press [1888] 1978.

Hume, David, "Of the Original Contract," in Eugene Miller (ed.), *David Hume - Essays: Moral, Political, and Literary*, Indianapolis: LibertyClassics, 1987, pt. II, Essay XII, 465-487.

Hume, David, "An Enquiry Concerning the Principles of Morals," In L.A. Selby-Bigge and P.H. Nidditch (eds.), *Enquiries Concerning Human Understanding and Concerning the Principles of Morals*, 3rd. edn. Oxford: Clarendon Press, 1989.

Kahneman, Daniel, and Amos Tversky, "Rational Choice and the Framing of Decisions," in Karen Cook and Margaret Levi (eds.) *The Limits of Rationality*, Chicago: University of Chicago Press, 1990, 60-89.

Kant, Immanuel, *Grounding for the Metaphysics of Morals*, James W. Ellington (trans.) Indianapolis: Hackett Publishing Co. [1785] 1981.

Latané, B. and J.M. Darley, *The Unresponsive Bystander: Why Doesn't He Help?* Englewood Cliffs: Prentice-Hall, 1970.

Lerner, Melvin J. and D.J. Miller, "Just World Research and the Attribution Process: Looking Back and Ahead," *Psychological Bulletin*, 85, 5, 1978, 1030-1051.

Lewis, David, "Prisoners' Dilemma is a Newcomb Problem," *Philosophy and Public Affairs*, 8, 3, 1979, 235-240.

Locke, John, *A Letter on Toleration,* Raymond Klibansky (ed.), J.W. Gough (trans.), London: Oxford University Press, 1968.

Locke, John, *The Second Treatise on Civil Government*, Buffalo: Prometheus Books, 1986 [Originally published as *Two Treatises of Government,* 1690]

Loomes, Graham and Robert Sugden, "Regret Theory: An Alternative Theory of Rational choice under Uncertainty," *Economic Journal*, 92, 805-824.

Luce, R. Duncan, and Howard Raiffa, *Games and Decisions: Introduction and Critical Survey*, New York: Dover [1957] 1985.

MacIntosh, Duncan, "Persons and the Satisfaction of Preferences: Problems in the Rational Kinematics of Values," *The Journal of Philosophy*, 40, 4, 1993, 163-180.

MacIntosh, Duncan, "Preferences's Progress: Rational Self-Alteration and the Rationality of Morality," *Dialogue*, 30, 1991, 3-32.

MacIntyre, Alisdair, *Whose Justice? Which Rationality?* Notre Dame, IN: Notre Dame University Press, 1987.

MacIntyre, Alisdair, *After Virtue: A Study in Moral Philosophy*, Notre Dame, IN: Notre Dame University Press, 1984.

Mackie, J.L., *Ethics: Inventing Right and Wrong*, Hammondsworth, Middlesex: Penguin, 1977.

McClennen, Edward, "Prisoner's Dilemma and Resolute Choice," in R. Campbell and L. Sowden (eds.) *Paradoxes of Rationality and Cooperation: Prisoner's Dilemma and Newcomb's Problem*, Vancouver: University of British Colombia Press, 1985, 94-104.

McClennen, Edward, "Constrained Maximization and Resolute Choice," *Social Philosophy & Policy*, 5, 2, 1988, 95-118.

Meyers, Diana, *Self, Society, and Personal Choice*, New York: Columbia University Press, 1989.

Mill, John Stuart, *Utilitarianism*, New York: Prometheus Books [London, 1863], 1987.

Mill, John Stuart, *On Liberty*, Elizabeth Rappaport (ed.), Hackett Publishing Co. [1859] 1978.

Moore, G.E., *Prinicipia Ethica*, Cambridge: Cambridge University Press, 1986.

Moore, Margaret, "Political Liberalism and Cultural Diversity." Unpublished manuscript, 1995.

Morris, Christopher, "Moral Standing and Rational-Choice Contractarianism" in Peter Vallentyne (ed.), *Contractarianism and Rational Choice: Essays on David Gauthier's "Morals by Agreement,*" New York: Cambridge University Press, 1991, ch. 6, 76-96.

Narveson, Jan, "Libertarianism, Postlibertarianism, and the Welfare State: Reply to Friedman," *Critical Review*, 6, 1, 1992, 45-82.

Narveson, Jan, "Gauthier on Distributive Justice and the Natural Baseline," in Peter Vallentyne (ed.), *Contractarianism and Rational Choice: Essays on David Gauthier's "Morals by Agreement,*" New York: Cambridge University Press, 1991, ch. 9, 127-148.

Narveson, Jan, *The Libertarian Idea*, Philadelphia: Temple University Press, 1988.

Nielsen, Kai, "Why Should I be Moral? Revisited," *American Philosophical Quarterly*, 21, 1, 1984, 81-91.

Nielsen, Kai, "Capitalism, Socialism, and Justice," in T. Regan and D. Van DeVeere (eds.) *And Justice for All*, Totowa, N.J.: Rowman & Allenheld, 1982, 264-286.

Nozick, Robert, *Anarchy, State, and Utopia*, New York: Basic Books, 1974.

Oldenquist, Andrew, "The Origins of Morality: An Essay in Philosophical Anthropology," *Social Philosophy & Policy*, 8, 1, 1990, 121-140.

Plato, *Republic*, Paul Shorey (trans.) in E. Hamilton and H. Cairns (eds.) *The Collected Dialogues of Plato*, Princeton: Princeton University Press, 1961.

Rawls, John, *A Theory of Justice*, Cambridge, Massachusetts: The Belknap Press of Harvard University Press, 1971.

Rawls, John, *Political Liberalism*, New York: Columbia University Press, 1993.

Rosenberg, Alexander, "The Biological Justification of Ethics: A Best Case Scenario," *Social Philosophy & Policy*, 8, 1, 1990, 86-101.

Rousseau, Jean Jaques, *The Social Contract*, Maurice Cranston (trans.) New York: Penguin Books [1762] 1982.

Runsiman, W.G., *Relative Deprivation and Social Justice: A Study of Attitudes to Social Inequality in Twentieth Century England*, Berkeley: University of California Press, 1966.

Russell, Bertrand, *The Problems of Philosophy*, New York: Oxford University Press, 1969.

Sandel, Michael, *Liberalism and the Limits of Justice*, New York: Cambridge University Press, 1982.

Sayre-McCord, Geoffrey, "Deceptions and Reasons to be Moral," *American Philosophical Quarterly*, 26, 1989, 113-122. Reprinted in Peter Vallentyne (ed.), *Contractarianism and Rational Choice: Essays on David Gauthier's "Morals by Agreement,*" New York: Cambridge University Press, 1991, ch. 12, 181-195.

Schmidtz, David, *Rational Choice and Moral Agency*, unpublished manuscript, 1994.

Schmidtz, David, "Rationality Within Reason," *The Journal of Philosophy*, 89, 9, 1992, 445-466.

Schmidtz, David, *The Limits of Government*, Boulder, Colorado: Westview Press, 1991.

Schopenhauer, Arthur, *On the Basis of Ethics*, E. F. J. Payne (trans.) New York: Liberal Arts Press [1840] 1965.

Schumm, George, "Transitivity, Preference, and Indifference," *Philosophical Studies*, 52, 1987, 435-437.

Smart, J.J.C., *Philosophy and Scientific Realism*, London: Routledge and Kegan Paul, 1963.

Smith, Holly, "Deriving Morality from Rationality," in Peter Vallentyne (ed.), *Contractarianism and Rational Choice: Essays on David Gauthier's "Morals by Agreement,*" New York: Cambridge University Press, 1991, ch. 14, 229-253.

Snyder, Mark, Elizabeth Decker Tanke, and Ellen Berscheid, "Social Perception and Interpersonal Behaviour: On the Self-Fulfilling Nature of Social Stereotypes," *Journal of Personality and Social Psychology*, 35, 1977, 656-666.

Sobel, Jordan Howard, "Cyclical Preferences and World Bayesianism," Paper read at the University of Waterloo, October, 1994.

Stevenson, Charles L., "The Emotive Meaning of Ethical Terms," *Mind*, 46, 1937.

Superson, Anita, "The Self-Interest Based Contractarian Response to the Why-Be-Moral Sceptic," *The Southern Journal of Philosophy*, 28, 3, 1990, 427-447.

Taylor, A. E., "The Ethical Doctrines of Hobbes," *Philosophy*, XIII, 1938, 406-424.

Taylor, Charles, *Sources of the Self: The Making of Modern Identity*, Cambridge: Harvard University Press, 1989.

Taylor, Charles, *Human Agency and Language*, Cambridge: Cambridge University Press, 1985.

Taylor, Charles, "Responsibility for Self," in A. Rorty (ed.), *The Identities of Persons*, Berkeley, University of California Press, 1976.

Trivers, Robert, "The Evolution of Reciprocal Altruism," *Quarterly Review of Biology*, 1971, 46, 1, 35-57.

Tversky, A. and D. Kahneman, "Judgment Under Uncertainty: Heuristics and Biases," *Sciences*, 185, 1124-1131.

von Neumann, John, and Oskar Morgenstern, *Theory of Games and Economic Behaviour*, Princeton: Princeton University Press, 1944.

Williams, Bernard, *Ethics and the Limits of Philosophy*, London, Fontana Press, 1985.