University of Alberta

David Gauthier's Moral Contractarianism and
the Problem of Secession

by

Edwin Ekwevukugbe Etieyibo

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Philosophy

Fall 2009
Edmonton, Alberta

Examining Committee

Dr. Wes Cooper, Philosophy

Dr. Adam Morton, Philosophy

Dr. Jennifer Welchman, Philosophy

Dr. Don Carmichael, Political Science

Dr. Jan Narveson, Philosophy, University of Waterloo

Dedication

To the ***"Ancient of Days," my Parents*****,**

And

TO THE MEMORY OF

Professor Campbell S. Momoh

Abstract

This thesis proposes a reading of David Gauthier's moral contractarianism (hereinafter Mb(CM)A) that demonstrates how cooperation can be rational in situations where expected utilities (EU) are stacked too high against cooperation. The dissertation critically examines Mb(CM)A and contends that it breaks down in the test of application, i.e. the problem of secession because of the conception of rationality it appeals to. Mb(CM)A identifies rationality with utility-maximization, where utility is the measure of considered coherent preferences about outcomes. Mb(CM)A links morality to reason, and reason to practical rationality, and practical rationality to interest, which it identifies with individual utility. On this view, an action (or a disposition) is rational if that action (or disposition) maximizes an agent's EU. This conception of rationality the essay claims is both naïve and misleading because it does not take into account an agent's considered preference for the acts that are available, in addition to the EU of those acts. Therefore, the thesis argues that Mb(CM)A's account of rationality be abandoned in favor of a decision-value/symbolic utility's or morals by decision-value agreement's conception of practical rationality. Morals by decision-value agreement (henceforth Mb(DV)A), the dissertation claims, handles serious problems, like the problem of secession in ways that Mb(CM)A cannot. Mb(CM)A breaks down in the test of application because when applied to the problem of secession, it suggests a single-tracked silver bullet solution. Specifically, it tracks only EU-reasons and claims that insofar as cooperation does not maximize the EU of better-off agents, it is not rational for them to cooperate with or support those

that are less well-off. By contrast, Mb(DV)A offers a multi-tracked framework for solutions to the problem, namely: it factors in an agent's considered preference for the acts that are available, in addition to EU of those acts. It is the argument of the thesis that when EU is stacked too high against cooperation, it may or may not be rational for an agent to cooperate, depending on which way symbolic utility (SU) for that agent points toward. If SU points in the direction of secession, then it is DV-rational for an agent not to cooperate, but if SU points toward non-secession, then it is DV-rational for that agent to cooperate.

Acknowledgements

This dissertation would not have been possible without the invaluable support of many people. My supreme gratitude goes to Professor Wesley Cooper, my supervisor, who made many sacrifices and volunteered much of his personal time to read over all the various drafts of the thesis. His positive comments, criticisms, and our discussions on Gauthier's moral views, contract and decision theories at Terwillegar Park while we walked Billy (Wes' dog) were very fruitful and deeply valuable. They enabled me eliminate numerous inconsistencies and obscurities, to formulate my ideas more sharply and with more clarity, and to strengthen my arguments at many points. On many fronts, Wes has been through and through a great source of support and encouragement.

My appreciation also goes to my co-supervisor, Professor Adam Morton and the other member of my supervisory thesis committee, Dr. Jennifer Welchman for their valuable time and feedbacks on the drafts of my work. I would like to thank as well Dr. David Kahane who once served as my co-supervisor before moving to the political science department. The many discussions I had with him helped me clarify and hone some of my ideas on Gauthier, secession, and on moral and political theory. I like to also thank the internal-external reader, Dr. Don Carmichael (Political Science) and the external reader, Professor Jan Narveson (University of Waterloo) for agreeing to be part of my examining committee. I particularly thank Don for his line of questioning during my candidacy exam. His thoughtful questions were provocative and they helped me fine-tune some of the fundamental ideas in the dissertation.

I am also indebted to the department chair, Dr. Bruce Hunter for his teaching support, and to Drs. Robert Burch and Glenn Griener for their support and constant encouragements. Glenn's sincere interest in my project helped me focused intensely on bringing the dissertation to fruition. Many thanks as well to Dr. Jayson MacLean (for copy-editing most of the thesis), to Anita Theroux and Wendy Minns for their steadfast departmental support, and to these friends, Dr. Paul Boaheng, Chris Sanford, Dr. Issac Amponsah, Frank Brai, Stephajn and Megan Saunter, Fatai Asodun, Ayo Adejumobi, Niyi Wahab and the Chukwuma brothers (Greg, John and Patrick). I am exceedingly grateful for your many kind support and encouragements.

I am most certainly grateful to my mom and siblings especially Barrister Jonathan Etieyibo, my cousins, Ovwata Ekhator and Avwega Omiegbe. Last but not the least I am thankful to my significant other, Maricel. Without their constant prayers, love, understanding and unwavering support I would not have been able to complete this project. I am extremely thankful to you all!

Table of Contents

List of Figures

Introduction

0.1 The Problem

David Gauthier's moral contractarianism is a theory of rational morality. It is an approach to moral theorizing and a recent and remarkably accomplished addition to the contractarian scholarship on morality and rationality. Morals by constrained maximization agreement (Mb(CM)A) is widely regarded by many scholars and commentators as the most systematic, sophisticated and rigorous in the social contract tradition. The technical and expert virtuosity that Gauthier displays in fashioning Mb(CM)A places social contract theory on a solid pedestal. Social contract theory flies higher with Gauthier's brand of contractarianism because he artfully and expertly brings together so many themes that are crucial to ethical, political and rational choice theories into a systematic whole, namely: the themes of consent and agreement, mutual advantage, preference and utility, cooperation, and maximization and optimization.

Gauthier develops a conception of practical rationality, one that gyrates on the wheel of rational choice theory, and then proposes a moral theory that he says is compatible with that conception of rationality. Mb(CM)A identifies rationality with the maximization of utility, where utility is the measure of considered coherent preferences about outcomes. On this view, an action (or a disposition) is rational if that action (or disposition) maximizes an agent's expected utility. Specifically, Mb(CM)A identifies rationality with utility maximization at the level of dispositions to choose, where those dispositions favor cooperation and allows an agent to maximize utility in some situation, given the strategies of those that he or

she interacts with. An agent chooses a disposition if and only if that agent expects to do better or maximize expected utility holding such a disposition than any alternative disposition.

As rigorous and systematic as Mb(CM)A is, its conception of rationality is naïve, narrow and misleading. Its identification of rationality with the maximization of expected utility precludes it from taking into account moral reasons or values that play a significant role in choice contexts, strategic and parametric. Specifically, it prevents it from considering an agent's preference or aversion for the acts that are available, in addition to the possible outcomes of those acts. Hence, when applied to the problem of secession it breaks down, namely, it offers a single-tracked silver bullet solution to a problem that requires a multi-tracked framework for solutions.

The problem of secession is a problem concerning what ought to be done to previously better-off, endowed and productive members of society who have become unproductive. Mb(CM)A's solution to this problem is suggested by its conception of rationality, according to which it is not rational for better-off members to cooperate with less well-off members. If the condition for accepting moral constraints on our behavior is that they satisfy mutual advantage and if what it means to satisfy mutual advantage is that those constraints advance our rational self-interest (i.e. maximize EU), then rational morality requires that we exclude from the contract or from any scheme of cooperation those unable to contribute to the cooperative surplus of that scheme. Simply put, because rationality requires that agents maximize EU it is not in the interest of better-off agents, according to

Mb(CM)A, to interact with less well-off agents in situations where EU is stacked too high against cooperation.

0.2 The Solution

This dissertation suggests a reading of Mb(CM)A that demonstrates how cooperation can be rational in situations where expected utilities are stacked too high against cooperation. This proposed reading replaces Mb(CM)A's conception of rationality with a decision-value/symbolic utility (DV/SU) or morals by decision-value agreement's conception of rationality. The replacement accepts the fundamental assumption of the sub-theory of constrained maximization insofar as rational or moral constraints are necessarily part of the fabric of the values that an agent brings to agreement and to cooperation.

Morals by decision-value agreement (Mb(DV)A) dissolves the problem of secession in ways that Mb(CM)A cannot because it takes into account an agent's considered preference or aversion for the acts that are available, in addition to their possible outcomes. Mb(DV)A's solution to the problem of secession is not a single-tracked silver bullet solution; it is a multi-tracked framework for solutions. Mb(DV)A claims that the act that an agent chooses in situations of secession is not informed by expected utility calculations but by DV calculations. It argues that whether better-off members choose to cooperate with or support less well-off members of society depends on whether the acts of cooperation or secession are intrinsically positively valenced for them, that is what values or weight they attach to the acts of cooperation or secession, in addition to the possible outcomes of

those acts. Specifically, the multi-tracked framework for solutions argues that it is DV-rational for better-off members to cooperate with less well-off members when symbolic utility points in the direction of cooperation and DV-rational for them not to cooperate when symbolic utility points toward secession.

0.3 The Thesis

In this section, I will be laying out the chapters that make up the thesis. Before I do that I think it might be helpful to provide a general outline of the main thrust of the dissertation. As a theory of rational morality, Mb(CM)A takes self-interest to be basic. Mb(CM)A is fundamentally an expected utility view of morality because it identifies rationality with rational self-interest which it identifies with the maximization of expected utility. Morality, according to Mb(CM)A, consists in certain types of cooperative behavior, i.e. those types of behavior that are mutually beneficial for self-interested agents to participate in. In other words, it is rational and *ipso facto* moral for an agent to participate in a scheme of cooperation only if that scheme speaks to the coherent considered preferences about outcomes of that agent or promotes his or her self-interest.

This dissertation critically questions this view of rational morality. It argues that it is rational and *ipso facto* moral for an agent to participate in a scheme of cooperation even if that scheme does not maximize that agent's coherent considered preferences about outcomes *provided*, of course, the scheme maximizes that agent's considered preferences for the acts that are available. In short, the

essay expands on the notion of self-interest that is basic to Mb(CM)A's conception of rationality.

There are at least three ways we can understand self-interest. Figure 0.0 illustrates and explains this.

***Figure 0.3a*: Three Views of Self-Interest**

1. *Unrefined Self-interest* (David Braybrooke): an individual's interest *qua* advantage is understood in terms of everyone's needs.

2. *Refined self-interest* (Gauthier/Mb(CM)A): an individual's interest *qua* advantage is understood in terms of that individual's preference *qua* expected utility.

3. *Super-refined self-interest* (DV/Mb(DV)A): an individual's interest *qua* advantage is understood in terms of that individual's preference *qua* decision-value.

Explanation:

1. *Unrefined self-interest* (David Braybrooke): needs are basic to interest. Needs trump preferences, i.e. are lexically prior to preferences. If there is a contract, then the contract individuals enter into is a *need-base contract* that satisfies needs before preferences. Simply put, self-interest is not identical with the maximization of EU, but with the satisfaction of needs.

2. *Refined self-interest* (Gauthier/Mb(CM)A): preference is basic to interest. Preference is satisfied when we maximize expected utility. The contract individuals enter into is a *preference-base contract* that satisfies preferences, but *considered coherent preferences about outcomes*, i.e. EU.

3. *Super-refined self-interest* (DV/Mb(DV)A): preference is basic to interest. Preference is satisfied when we maximize decision-value. The contract individuals enter into is a *preference-base contract* that satisfies preferences, both *the considered coherent preferences about outcomes* (EU) and *the considered preference for the acts themselves*, i.e. SU (or the meaning or expressiveness of the acts).

My thesis claims that Mb(DV)A is superior and theoretically more fundamental than Mb(CM)A because its account of rationality appeals to super-refined self-interest. To bring out the intuitive force of this claim consider the application of Mb(CM)A and Mb(DV)A to parametric and strategic contexts.

(i) Parametric contexts:

These are contexts where an individual's choices are independent of the choices of others, i.e. an individual's choices are the sole variables in choice situations. Consider the following example. X finds a lost wallet. The wallet has enough money. If X takes the money in the wallet no one will find out. Is it rational or not for X to keep the wallet?

Mb(CM)A claims that it is rational for X to keep the wallet. Why? Because the act maximizes his or her expected utility. Mb(DV)A agrees with Mb(CM)A that keeping the wallet maximizes X's expected utility. It however, maintains that it may be rational (not EU-rational but DV-rational) for X not to keep the wallet. According to Mb(DV)A, it is rational for X to return the wallet if and only if the act (i.e. keeping the wallet) is positively intrinsically valenced for him or her. Simply stated, it is rational for X to keep the wallet if SU points toward that direction (factoring the EU of the acts of keeping the wallet and returning it).

(*ii) Strategic contexts:*

These are contexts where an individual's choices are not the sole variables in choice situations. That is to say, an individual's choices are partly dependent on

that individual's expectations of the choices of others, and vice versa. We can illustrate strategic contexts with the Prisoner Dilemma and the Problem of Secession.

First, *the Prisoner Dilemma*: X and Y enter into a contract to babysit for each other on weekends their favorite band comes to town. The choice of each is dependent on the expectation of the choice of the other. Let us represent this with the following matrix:

Figure 0.3b: **The Prisoner Dilemma with Matrix Showing Utilities**

		X	
		Babysit	Don't Babysit
Y	Babysit (cooperation)	25, 25	0, 35
	Don't Babysit (non-cooperation)	35, 0	15, 15

From the graph, non-cooperation strictly dominates cooperation. And according to the standard response, i.e. orthodox rational choice theory it is rational or in the interest of X and Y not to cooperate. But for X and Y not to cooperate is to fail to maximize their EU. Specifically, to choose the non-cooperative act is to choose a sub-optimal outcome. The cooperative act is the optimal outcome.

Mb(CM)A attempts to solve this problem by appealing to dispositions. Either of X and Y chooses appropriate dispositions, namely, each chooses those dispositions that favor the cooperative optimal-outcome. The case then is this: it is

rational for X to babysit for Y if she expects that (1) Y would babysit for her and (2) cooperation provides her optimal outcomes, i.e. maximizes EU. Since both X and Y are rational and since either aims to maximize EU, each would adopt dispositions that favor the cooperative optimal-outcome and each would rationally constrain his or her straightforward maximizing behavior (behavior that aims for individual utilities but which leads to sub-optimal outcomes). Hence, X will babysit for Y because X knows Y has formed the disposition to babysit for her, and Y will babysit for X because Y knows X has formed the disposition to babysit for him. If X knows Y has not formed the dispositions to cooperate, she would not babysit for him, and if Y knows X has not formed the dispositions to cooperate, he would not babysit for her. But since X gains from Y babysitting for her and since Y gains from X babysitting for him, they would form those dispositions that favor the cooperative optimal-outcome.

But dispositions are problematic, especially the view that they are accessible or transparent and have some 'quasi magical properties.' Appealing to dispositions does not seem to solve for Mb(CM)A the problems that confront decision-makers in strategic contexts. Even if we assume that by appealing to dispositions Mb(CM)A solves some or all of the problems in strategic contexts, it is clear that this is at the expense of making use of extra assumptions or auxiliary hypotheses or at the cost of sacrificing simplicity and 'multiplying entities and essences beyond necessity.'[1]

[1] The principle that essences or entities should not be multiplied unnecessarily is *Occam's razor* or *principle*, which is attributed to 14th century English logician and Franciscan friar, William of Ockham. The principle is also called the "Law of Parsimony." A popular application of the principle simply states that when there are two competing theories that make exactly the same

There are two reasons why Mb(DV)A is theoretically more fundamental than Mb(CM)A. Firstly, it fares better when applied to these same problems. Secondly, it dissolves the problems without appealing to the additional assumptions or 'essences' that Mb(CM)A appeals to. Mb(DV)A does not appeal to the 'quasi magical properties' of dispositions and all the other baggage that come with it. For Mb(DV)A *simpliciter*, it is rational for X to babysit for Y even if she expects that Y would not babysit for her, *provided* that the act of babysitting for Y is positively intrinsically valenced for her. If by jettisoning the 'quasi magical properties' of dispositions and all the other baggage that come with it makes Mb(DV)A a simpler and less messy theory, then this seems to be a clear case where the more plausible theory is the simpler and less messier one.

Second, *the Problem of Secession*: the problem of secession is a problem concerning what ought to be done to previously better-off, endowed and productive members of society who have become unproductive. If, according to Mb(CM)A, what rationality requires us to maximize is expected utility, then when expected utility is stacked too high against cooperation, it is rational for better-off agents not to cooperate with less well-off agents. There are two objections here.

The first is that Mb(CM)A offers a single-tracked silver bullet solution to a problem that requires a multi-tracked framework for solutions. The second objection comes from Braybrooke. According to this objection, Mb(CM)A fails the crucial test of morality, namely, the test of moral concern, according to which needs must be prioritized over preferences. Mb(CM)A fails the test of moral

---

predictions, the simpler one is the better. The normative force of the principle is that it requires a theory or an explanation of a phenomenon to make fewer assumptions and to eliminate those that make no difference in the observable predictions of the explanatory theory.

concern because it, Braybrooke claims, assumes and wrongly at that, that all that we can find in morality is put there by reason alone. Thus, for Braybrooke, it should be abandoned in favor of an affective-base/need-base account of morality if we are to have a moral theory that is appropriately suited for agents *qua* humans and to deal with moral progress.

To take care of these objections the thesis proposes that we modify Mb(CM)A's notion of self-interest, which essentially is a replacement of Mb(CM)A's account of rationality with a Mb(DV)A's conception of rationality. An Mb(DV)A's conception of rationality provides a multi-tracked framework for solutions to the problem of secession. A multi-tracked framework for solutions takes into account all moral reasons that play a significant role in an agent's choice. It recognizes that an agent may switch between different choices, solutions, or strategies, or acts depending on the expressiveness or meaning of the acts that are available (factoring as well the EU of those acts). If it is rational to maximize DV, then when EU is stacked too high against cooperation, it may or may not be rational for an agent to cooperate, depending on which way SU for that agent points toward. If SU points in the direction of secession, then it is DV-rational for an agent not to cooperate, but if SU points toward non-secession, then it is DV-rational for that agent to cooperate.

The thesis is divided into six chapters. The first chapter is a brief overview and survey of the social contract tradition. In this survey, I shall be examining the two distinct strains of social contract thought: contractarianism and contractualism. My focus in this chapter is on two themes that converge around the general

problem of rational compliance: the legitimacy of political government, and morality and utility-maximization. In examining the first theme, I will be examining how classical (17th and 18th century) social contract accounts cash out the relationship between the contract and morality or politics. And in examining the theme of morality and utility-maximization, I shall be examining the general problem posed by the freerider *qua* rational skeptic for cooperation and morality in general and Gauthier's moral contractarianism in particular. The chapter concludes with an outline of the problem of secession and Mb(CM)A's solution to it.

Chapter two examines Rawls' contribution to the social contract tradition, i.e. *Justice as Fairness*. My examination of Rawls' theory focuses on the relationship between the various assumptions—reflective equilibrium, 'the veil of ignorance,' the maximin rule, and 'a capacity for a sense of justice'—Rawls makes and the principles of justice that he says will be chosen by contractors in the 'original position.' I shall be limiting my consideration of *Justice as Fairness* (JaF, for short) to the contract framework, as Rawls sets it up. Specifically, I shall be examining the way he employs the resources of rational choice theory to calibrate the principles of justice that emerge from the contract in the 'original position' and why he thinks the principles are those that would be chosen in appropriately idealized and specified conditions. In doing this, I shall be highlighting what is mistaken about Rawls' social contract theory, and how his theory contrasts with Gauthier's, which I believe is theoretically more fundamental.

Mb(CM)A is theoretically more fundamental than JaF because it makes fewer assumptions—one of which is the proviso, which ground property rights

that, Gauthier argues, are crucial for the emergence of moral principles. Although moral contractarianism makes fewer assumptions, in particular doing so without the assumptions of reflective equilibrium, the veil of ignorance, and a capacity for a sense of justice, it has the same explanatory power as JaF, in the sense that it is able to explain why moral principles are rational and obligatory. The difference about making assumptions is, in particular, a difference about the extent to which Mb(CM)A and JaF rely on moral intuitions, namely, what Rawls calls 'our considered moral judgments' or 'beliefs.'

Furthermore, Mb(CM)A's characterization of an essentially just society is more determinate and realistic than the characterization that JaF offers. An essentially just society, for JaF, is one that satisfies mutual advantage, which is identified with the standpoint of the least advantaged member of society. For Mb(CM)A, an essentially just society is one that satisfies mutual advantage—explained by the considered coherent preferences of rational actors—and which converges and coheres with the standpoint of the Archimedean chooser. The Archimedean standpoint is a hypothetical vantage point from which an individual can affect some object or objectively perceive with totality the subject of inquiry. In moral theory, the Archimedean standpoint is that position an individual must occupy if that individual is to have the moral capacity to affect society, i.e. it is that vantage point that an individual must be in if that individual's "decisions are to have the moral force needed to govern the moral realm."[2]

[2] Gauthier, *Morals by Agreement*, Oxford, Oxford University Press, 1986, p. 233. *Morals by Agreement*, although refers to Gauthier's book also refers to his moral contractarian project. Henceforth, I would simply refer to *Morals by Agreement* as MbA. When I use the acronym MbA I

To index mutual advantage or the contract framework to the standpoint of both the individual and Archimedean choosers is to contour the principles or constraints of morality to reflect the interest of every rational chooser. Because moral principles reflect everyone's interest, Gauthier shows he values seriously the thesis of individualism. The thesis stipulates that we may neither collapse a person's conception of the good with those of others nor compel anyone to accept the principles of social relationships. On the other hand, to index mutual advantage or the contract framework to the standpoint of the least advantaged member of society is to shape the principles of justice to reflect the interests of the least advantaged group. Shaping the principles of justice to reflect the interests of the least advantaged chooser is to violate the thesis of individualism and to require that some people engage in activities for the benefits of others, i.e. to treat some people as means to the ends of others.

In chapter three, I shall examine Gauthier's moral contractarianism, which seeks to demonstrate that (1) constraints are fit for utility-seeking agents, and (2) constraints agreed to by free and rational persons advance or promote rational self-interest. Among other things, this chapter focuses on how Gauthier's approach to moral theorizing addresses the general problem of compliance. There are three core sub-theories or components in Mb(CM)A: (a) minimax relative concession and the bargaining problem; (b) constrained maximization and the problem of rational compliance; (c) the contract problem or the problem of specifying an appropriate natural baseline from which rational bargain is to proceed. I shall be examining the

---

am referring to the book *Morals by Agreement*, and when I use 'Morals by (Constrained Maximization) Agreement (Mb(CM)A) I am referring to Gauthier's moral contractarian project.

sub-theories in the order listed. Following my examination of the sub-theories, I will conclude by examining what makes the 'liberal individual,' in Gauthier's view, superior to the 'economic individual.' I shall focus on how the former's possession of "affective capacity for morality" contrasts both with the view of the economic individual and with the "capacity for an effective morality" that moral thinkers like David Hume defend. By defending the liberal individual's affective capacity for morality, Gauthier is able to make the case that moral contractarianism does not banish a vigorous view of participatory activities, notwithstanding the fact that it requires an agent to display a non-tuistic interest when interacting with others.

In chapter four, I shall examine broadly the test of application for Mb(CM)A: the problem of secession. My examination of this problem proceeds from my analysis of the scope of the contract problem. The problem of secession is a problem of what should be done with previously endowed or productive members of society. The scope of the contract problem is a problem of whether the contract should be a contract of sub-groups or groups, or national societies, or of a society of the human race. My focus here is to discuss the problem and Mb(CM)A's solution to it. I take up further, in this chapter, the issue of the capacity for an affective morality by examining Hume's theory of moral sentiments. Part of my reason for doing this is to see if an account of affective morality justifies Braybrooke's negative and positive theses: that Mb(CM)A or any social contract theory cannot solve the problem of secession, and that an account of moral sentiments is able to dissolve the problem of secession. Mb(CM)A's morality is the

morality of the liberal individual, namely, the morality of free affectivity. Because the emotions and feelings of Mb(CM)A's agents are only engaged by the demands of rationally based constraints, its morality is different from the morality of a theory of moral sentiments.

In chapter five, I shall examine how Mb(DV)A revises Mb(CM)A's account of practical rationality and how this revision provides a multi-tracked framework for solutions to the problem of secession. Mb(DV)A's revision takes into account the possibility of an agent's considered preference or aversion for acts, in addition to the EU of those acts. EU is outcome-sensitive because it is about the possible consequences of actions. In addition to EU, there are utilities that are about what the actions symbolize, or express, or mean. We may call these utilities action-sensitive to distinguish them from outcome-sensitive utilities. Mb(CM)A subscribes to an EU-view of rationality and morality; it is EU-focused or sensitive, i.e. it identifies rationality with the maximization of utility, where utility is the measure of considered coherent preferences about outcomes. Mb(DV)A, on the other hand, does not subscribe to an EU-view of rationality and morality. Mb(DV)A is not EU-focused, rather it is both EU and SU focused; it identifies rationality with the maximization of decision-value, i.e. an agent's considered preference for the acts that are available (SU), in addition to the EU of those acts. I begin my discussion, in this chapter, by examining how a decision-value account resolves long-standing paradoxes, such as Newcomb's Problem and the Prisoner Dilemma, after which I examine how Mb(DV)A relates to the general structure of desire-based and value-based accounts of practical reasons. I then show how

Mb(DV)A handles the problem of secession and the sense in which it provides a middle way between the strict liberal individualism of Mb(CM)A and the thoroughgoing or extreme communitarianism of Rousseau's social contract theory. I conclude by making the case that Mb(CM)A and accounts of affective moralities occupy different ends of the spectrum of silver-bullet accounts.

In the final chapter, I shall examine three critiques of Mb(DV)A. Foremost among these is the critique that although Mb(DV)A provides a framework for discriminating between situations that are DV-rational (when symbolic meanings and preferences are sufficiently appealed to) and situations that are DV-irrational (when symbolic meanings and preferences are sufficiently appealed to), it does not provide a framework for distinguishing between which symbolic meanings and preferences are good or desirable and which symbolic meanings and preferences are bad or undesirable. The two other critiques I will be examining are: should peoples' characters be shaped to have or not to have various values or symbolic meanings and preferences, or SU-reasons for being united with others, or for cooperating with others? And, does Mb(DV)A not violate the demand of an essentially just society as a cooperative venture for mutual advantage?

Chapter One

The Social Contract Tradition: A Brief Overview and Survey

Introduction

Social contract theory is one of a handful of truly most influential theories within moral and political theory in the history of the contemporary West. The social contract tradition is a very great one—going all the way back to Thomas Hobbes. Hobbes was the first thinker to give a complete exegesis and defense of social contract theory. That theory was resurrected in the 1970s and the decades after it, following the publication, in 1971, of John Rawls' highly influential book, *A Theory of Justice*. Rawls' views in *A Theory of Justice*[3] generated renewed philosophical interest in moral and political philosophy in general and in the social contract thoughts of Hobbes, John Locke, Jean-Jacques Rousseau and Immanuel Kant in particular.

Social contract theory is grounded in individual interest and it describes a broad class of theories that try to explain and justify morality or politics. It is the view that moral or political obligations[4] are derived from the contract or agreement made between rational persons to form societies.[5] The central idea underlying a

[3] Rawls, *A Theory of Justice*, Cambridge, Harvard University Press. Further reference to *A Theory of Justice* will be abbreviated as ToJ, and except indicated otherwise this will refer to the revised edition of 1999.

[4] I take morality to refer primarily to moral norms or principles that regulate moral behavior among individuals, and politics to refer primarily to principles that govern the relationship between individuals and civil or political society, i.e. the state or government. I shall sometimes be using the term 'principles of social relationships' to refer to both morality or moral obligations or principles and politics or political obligations.

[5] It is important to emphasize that although contract theorists specify the social contract in terms of agreement made between rational persons to form societies they specify the content of this agreement or what it is about differently. For example, Rousseau specifies it as a policy that works equally to the interest of all; Rawls, as principles that shape the basic structure or institutions of

social contract view is that morality, social and political practices and institutions are acceptable to fully rational persons if and only if they *could theoretically* or *in principle* agree to them, or just in case they could be *rationally justified* to everyone.

To begin my analysis of the social contract tradition it would be helpful to point out the two distinct strains of social contract thought: *contractarianism* and *contractualism*. Whereas contractarianism has its origin in Hobbes, contractualism has its roots in Rousseau and Kant. Hobbes' contractarian account is founded on mutual self-interest. Under contractarianism, individuals seek to maximize their own interests in a rational bargain with other individuals. Morality or politics it claims consist in certain types of cooperative behavior: those types of behavior that are mutually beneficial for self-interested agents to participate in.

Two main ideas are fundamental to contractarian thought: first, that rational agents are primarily self-interested and second, that a rational evaluation of the best strategy for maximizing their self-interest will lead them to accept the authority of political government or to act morally. Contractarianism, one must note, is a broad term that refers either to a political theory of the legitimacy of political authority or to a moral theory about the origin of moral norms. When it refers to the first contractarianism claims that state power must come from the consent of the governed, where the form and content of this consent derives from the idea of the contract. And when it refers to the second contractarianism holds that the normative force of morality or moral principles is derived from the idea of the

---

society for the purpose of achieving equality and fairness; Hobbes and Locke, as the commitment to give up some or all of one's rights to a political government; Gauthier, as the adoption of a rational disposition to cooperation or to be moral.

contract. The most well known advocate of this view is Gauthier, whose brand of contractarianism, 'Morals by Agreement' I will be examining in this essay.

Contractualism has its origin in Rousseau and Kant. Rousseau's social contract theory is based on the general will. Rousseau takes the general will as an expression of freedom, i.e. a policy that actualizes the individual's rational freedom. To claim that the general will is an expression of freedom is to claim that it is a policy that works in the interest of everyone or equally well for all concerned and that is adopted jointly by free and equal citizens. Like Rousseau, Kant's account of the social contract is founded on the idea of rational freedom. The aim of the social contract, he claims, is to seek principles that all rational persons would agree or subscribe to, under certain idealized conditions. In order to arrive at these principles and to reach agreement, Kant, abstracts away from many concrete and determinate features of the moral lives of rational beings.

In contrast to contractarianism, contractualism does not claim that individuals seek to maximize their own interests in a rational bargain with other individuals. Rather, it claims that individuals seek to pursue their own interests in ways that they can justify to others who also have their own interests to pursue. There are two related theses underlying a contractualist view: firstly, that rationality requires that we respect the equal moral status of persons and secondly, that rationality requires that morality or state power (politics) be rationally justified to each person. It is important to point out that contractualism cashes out the equal moral status of persons in terms of rational autonomous agency. Morality or politics it claims consist in what would result if we were to create obligatory

agreements from a point of view that respects everyone as rational autonomous person. The most prominent exponent of this view is Rawls who, like Kant, holds that the purpose of the contract is to seek principles to which representative members would agree under appropriately specified initial conditions.

Rawls' contractualism is political in the sense that it seeks to set the general social framework for a liberal society. Like Kant, Rawls abstracts away from many concrete and determinate features of our moral lives. In order to choose the principles that will govern the basic institutions of society, Rawls screens out many characteristics of his agents by placing them behind a 'veil of ignorance' in the 'original position.'[6] According to Rawls, we ought to comply with the principles of justice that would be rational for every person to choose, if we have to choose those principles without knowing anything about our particular characteristics and social and economic circumstances.

To the extent that social contract theory is grounded in interest and agreement and seeks primarily to justify the principles of social relationship as they affect us, it is faced with a number of fundamental questions: (1) "what is the nature and status of the principles of social relationships? (2) What are the process and mechanisms that give rise to them? (3) How extensive are they? (4) Given the limit that the principles impose on behavior, why should anyone accept them?"

[6] Rawls takes the 'original position' to be identical with the state of nature in the sense that it is a pre-political state. The state of nature is a state that lacks moral principles or political authority. The original position is a fair and impartial point of view; the point of view that allows agents to hypothetically reason impartially about fundamental principles of justice. The 'veil of ignorance' is a device that Rawls employs in the original position to ensure impartiality of judgment about the principles of justice. In the original position, agents placed behind the veil of ignorance are deprived of all knowledge of their personal characteristics, social and economic circumstances. I examine the original position and the veil of ignorance in chapter two.

We can evaluate the significance of the principles of social relationships through an examination of the nature of human interactions vis-à-vis human needs. Our needs and interests are diverse and they define, for the most part, the reasons why we seek to interact with others. The role that our needs and interests play in human interactions is obvious in that in any interaction we seek the best situation that satisfies them. If we view the satisfaction of our needs and interests as 'benefits' and what we give up to get these benefits as 'costs', then we might say that the most beneficial type of interaction is one where marginal benefits are symmetrical to marginal costs. To this extent, marginal costs and benefits in any interaction to any person necessarily bear one to three possible relationships:[7]

(i) Marginal benefits are less than Marginal costs (Mb < Mc)

(ii) Marginal benefit are greater than Marginal costs (Mb > Mc)

(iii) Marginal benefits are equal to Marginal costs (Mb = Mc)

Before proceeding any further let me define the sense in which I will be using these expressions: (a) marginal costs and marginal benefits in natural interactions and (b) marginal costs and marginal benefits in social interactions. Whereas natural interactions are interactions where the principles of social relationships are not operational, social interactions are interactions regulated by the principles of social relationships. I will use marginal costs in natural interactions to mean broadly the additional costs that individuals incur in the absence of the principles of social relationships, i.e. the added price individuals pay

[7] Gauthier, "Three against Justice: the Foole, the Sensible Knave, and the Lydian Shepherd," in *Moral Dealing: Contract, Ethics and Reason*, Ithaca and London, Cornell University Press, 1990, p.132.

when their 'liberties' are violated by others. And I will use marginal benefits in natural interactions to refer broadly to the extra gains individuals receive from violating the 'liberties' of others. I will be using marginal costs in social interactions to mean broadly the added costs individuals internalize for cooperating with others, i.e. the extra price individuals pay for complying with the principles of social relationships. And when I will use the expression marginal benefits in social interactions I will use it to refer broadly to the additional benefits that individuals receive from complying with the principles.

What do I mean by benefits and costs? In the context of my discussion, benefits and costs have a broad usage. I will follow the general outline of the different ways that they have been specified by social contract theorists. Under social interactions, benefits have been specified as follows: as self-preservation (Hobbes); as the realization and preservation of freedom or our equal moral status (Rousseau, Kant, and Rawls); as the preservation of property (Locke); as the maximization of expected utility (Gauthier). And the costs under social interactions have been specified as follows: the liberties and rights—natural or social—that are given up either to a political authority or to establish morality; the constraints on egoism or self-seeking behavior.

In the first scheme or relationship, marginal benefits are less than marginal costs and this reflects the presence of negative externality or external inefficiency. In scheme (ii), marginal benefits are greater than marginal costs, which points to the presence of positive externality. The excess of marginal benefits over marginal costs in this scheme we may suppose arises, according to Gauthier, because the

interaction "is not purely instrumental—interaction among persons who take an interest in one another's interests."[8] If marginal benefits are equal to marginal costs, as in the third relationship, interaction is mutually advantageous, and positive externality and negative externality cancel each other.[9] Social contract theory seeks to demonstrate how the third scheme is possible.

Because the first and second schemes are not mutually advantageous, they are considered suboptimal. When there is a surplus in marginal costs over marginal benefits costs are imposed on others without any corresponding benefits to them. In essence, when marginal benefits are less than marginal costs "freeriders" are given a free pass. Freeriders are those who bear less than a fair share of the costs of production of a benefit. In its less severe form, a freerider contributes only a little to the costs of producing benefits, but gets more or the same amount of benefits as others. In its more severe form, the freerider accepts the principles of social relationships but breaks them when it is in his interest to do so.[10]

Freeriding is a problem for any social interaction or scheme of cooperation. And the 'freerider problem' is a question of how to prevent any form of freeriding from taking place. In dealing with the freerider problem, it is important that the principles of social relationships that regulate both individual and collective actions be rationally acceptable to every person. The principles are meant to distribute the costs of interaction fairly among those that benefit from such interaction. The problem is not that the freerider does not accept the principles of social relationships. Like everyone else, he accepts them. But the problem is that he is

[8] Ibid, p.132.
[9] Ibid, p.132.
[10] My use of his, he or him throughout my discussion of the freerider has no sexist undertone.

prepared to break them if it would be to his benefit to do so. So in addition to demonstrating how the principles of social relationships arise and the role they play in regulating individual and collective actions, it is important that social contract theorists provide the freerider reason or reasons as to why he should not break them.

Now, one must point out that although contractualism and contractarianism for that matter are concerned with the role that the principles of social relationships play in regulating individual and collective actions, it is the latter that sets out to provide a systematic response to the freerider problem. This is not to say that contractualism does not recognize the problem freeriding creates for social interactions or schemes of cooperation. Given the fundamental claims of contractarianism, it is not unexpected that contractarianism and not contractualism should set out to provide a more systematic response to the freerider problem. Since what individuals seek under contractarianism is the maximization of their own interests in a bargain with other individuals, it, and not contractualism has the additional burden of providing the freerider reason or reasons as to why he should not break the principles of social relationships, given that it seems to maximize his interest to do so.

There are legitimate reasons for contractarianism to be worried about the freerider problem. The reasons are generally summed up as 'the problem of instability.' The instability problem is the problem of avoiding the complete breakdown of schemes of cooperation, due to the lack of support from those that benefit from those schemes. Of course, there are different degrees of instability and

it is beyond the scope of this thesis to consider them. The general thrust of the problem of instability is this: because freeriding places the burden of producing benefits on some, those who produce the benefits from which the freerider gains may find it unacceptable to internalize the costs necessary to produce them. The idea is that a scheme of cooperation that breeds and encourages freeriding is not mutually advantageous. That scheme allows others to be used as mere means, i.e. to benefit at the expense of others. A freeriding behavior is therefore, a threat to cooperative schemes because it smothers and blunts the motivation of those that provide for their maintenance. The point about the problem of instability as it concerns freeriding is straightforward, and it is this: there would be no benefits to share from when those who produce benefits refuse to take up the costs necessary to produce them or when those who support and maintain a scheme of cooperation decide not to contribute to their maintenance.

We have an optimal situation when marginal benefits are equal to marginal costs as in the third scheme above. This is because the costs imposed on any individual are proportional to the benefits that the individual receives. In this scheme there is no freeriding. Cooperation requires that individuals internalize some costs, i.e. accept some obligations. If we say that agents in general are interested in their own good or wellbeing, then we would expect them to internalize the costs of maintaining a scheme of cooperation that provides them benefits.[11] Hence, it seems right to say that to get an individual to voluntarily accept and

[11] My discussion here is a general discussion of the costs and benefits of social interactions or cooperation. It anticipates the problem that arises for an individual who has to choose between cooperation and noncooperation. It does not anticipate the problem that arises for an individual who has to choose among different schemes of cooperation.

maintain social institutions and practices the marginal benefits, for that individual, *must* not be less than the marginal costs that that individual pays for their maintenance. Given that it is not in an individual's interest to produce benefits for others or to invest in a scheme of cooperation more than he or she benefits, freeriding presents a threat to those that support social institutions and practices, to cooperation, and to the foundation of morality.

But why should the freerider care? After all his interests are served by his straightforward maximizing behavior. He is aware that he reaps benefits without incurring the costs necessary to produce them, but why should that be of consequence to him? Why should it matter to him that social institutions and practices, cooperation or the foundation of morality are undermined by his self-serving straightforward maximizing behavior? He has no reason or motivation to constrain his behavior even though it reduces the benefits of others or places the burden of producing them on a few. Provided that he benefits from such conduct, he feels good about himself. He might be acting 'immorally' but certainly not 'irrationally.'

The rationality of the freerider is fundamentally the rationality of the 'rational skeptic', which we might identify with the general problem of rational compliance. The rational skeptic wants us to provide him a reason or reasons as to why he should cooperate with others or why he should be moral in situations where acting 'immorally' would benefit him. If we associate the rational skeptic with Hobbes' 'Foole,' David Hume's 'Sensible Knave,' and the 'Lydian Shepherd'—all of whom are not averse to engaging in behaviors that threaten cooperation and

morality—then the general problem of rational compliance is essentially the Prisoner's Dilemma, which will be discussed later.[12] The rational skeptic neither cares tuppence nor an iota about breaking rational agreements or the principles of social relationships as long as his interests, whatever these happen to be, are satisfied. Whatever concern he has for social institutions and practices and morality is determined strictly by his self-serving interest.

For the rational skeptic, it is both reasonable and acceptable to eke out extra benefits whenever the situation or one's interest calls for it. It is acceptable to improve one's position if that means taking advantage of the 'good behavior' of others. It is reasonable to reap additional benefits for oneself if that means not honoring the agreements one has previously made. For the rational skeptic, reason and interest not only part ways, but direct rational agents to conflicting actions. Reason and interest direct us to accept restrictions on our liberties, which reason and interest then seem to subvert. Reason, on the one hand, tells us it is in our best interest to voluntarily accept the principles of social relationships, and on the other hand, tells us to break them, when doing so is conducive and congenial to our self-interest.

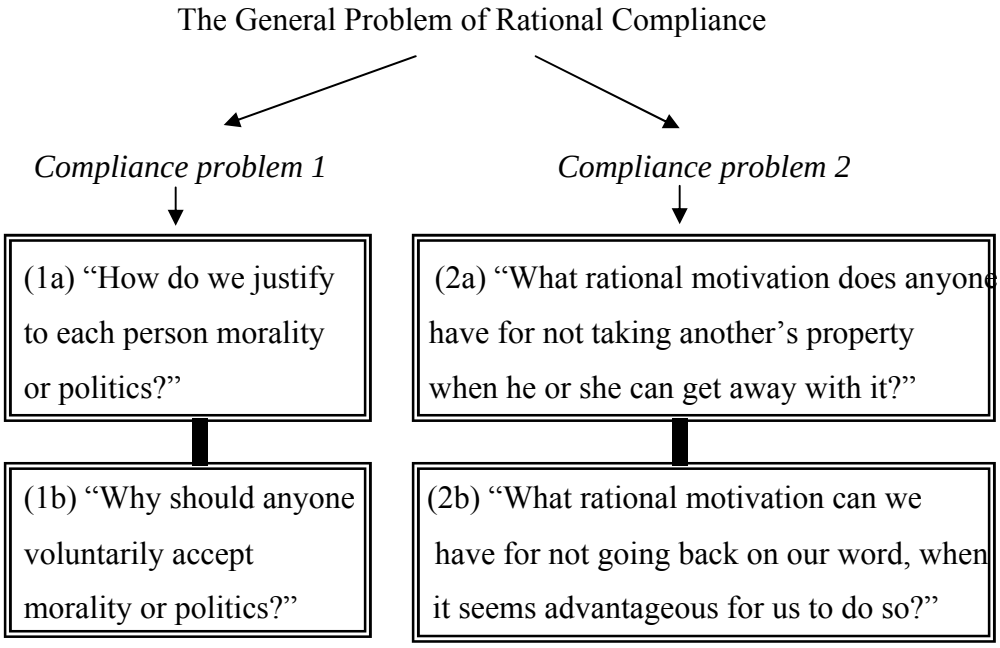
Note that the general problem of compliance arises because of the seeming conflict between self-interest and reason. There are two parts to this problem, which are related to each other. The first part of the problem—let us call this

[12] In MbA, Gauthier defines the problem of stability as "the willingness of others voluntarily to accept social institutions and practices," p.344. In his discussion of the problem of stability, he suggests that the unwillingness of people to voluntarily accept social institutions and practices is a fundamental problem for the contract. See also pp.129-149 of *Moral Dealing: Contract, Ethics and Reason* for Gauthier's discussion of the various ways the Foole, Sensible Knave, and the Lydian Shepherd try to undermine morality, rationality and justice.

'compliance problem 1'—demands that social contract theory demonstrate that the principles of social relationships or a scheme of cooperation benefit everyone. The second part of the problem—which we shall call 'compliance problem 2'—demands that social contract theory provide a reason or reasons as to why an individual should comply with the principles of social relationships in situations where it seems not to be in that individual's interest to do so. In other words, compliance problem 2 requires that social contract theory give reasons as to why a person should not violate the principles and requirements of a scheme of cooperation whenever the opportunity to make great gains arise or steal from others if that person is absolutely sure to get away with it or if the possibility of detection is relatively low.

It is important to point out that accounts of contractualism do not primarily seek to provide a systematic response or to solve the general problem of compliance. However, the fact that they claim that the contract engenders a scheme of cooperation that works to the benefit of everyone—whether by promoting rational autonomous agency or by respecting our equal moral importance as rational autonomous agents—one can say that they are indirectly concerned with the first part of the compliance problem. If we frame the fundamental questions facing social contract theory, i.e. the four questions I raised earlier in terms of the general problem of rational compliance, we would have the following questions for compliance problems 1 and 2:

Figure 1.0: **Two Parts of the Compliance Problem**



In the sections that follow, I will briefly examine the views of social contract theory on the relationship between the contract and morality or politics. I will examine in section 1.1 the justification of political government or the state by classical (17^{th} and 18^{th} century) accounts of social contract theory. In subsections 1.1.1 and 1.1.2, I shall respectively look at how classical accounts of contractualism (Rousseau and Kant) and classical contractarian accounts of Hobbes and Locke justify the authority of political government. In section 1.2, I shall introduce Gauthier's moral contractarianism. This will be followed by a brief introduction and outline, in section 1.3, of the problem of secession, Mb(CM)A's solution to it and the multi-tracked framework for solutions that I shall be defending in this thesis.

1.1 Classical Social Contract Theory and Legitimate Political Government

In view of the wide-ranging impact that political government has on our individual lives, the question why should we accept its authority seems an appropriate one to ask. The question of the justification of the authority of the state, as it relates to classical social contract theory, can be broken down into the following two theses: (1) political government derives its authority from the positive outcomes it makes possible, (2) those that it has authority over must consent to it. Classical social contract accounts claim to be able to argue for both theses. The argument that these accounts provide for the legitimacy of political government is essentially an argument for the principles of social relationships, which bring together people under the umbrella of one common political or moral authority.

Call the first thesis 'a service conception' view of political government or authority, according to which the justification of political government derives from how well it benefits the people. Now, 'a service conception' view of authority is part of the broad demand of mutual advantage. If we understand the demand of mutual advantage normatively, then it stipulates that any scheme of cooperation must work to the benefit of everyone. A scheme of cooperation works to everyone's benefit when the costs an individual pays for maintaining that scheme are not in excess of the benefits that he or she receives. Hence, an account argues for the first thesis when that account successfully demonstrates that what an individual gives up in order to be under the authority of political government is at least symmetrical to the benefits that such an individual receives.

The second thesis constitutes part of the principle or thesis of 'individualism.' This thesis is also called the thesis of 'separateness of persons.' The thesis of individualism stresses the paramount importance of individual independence and self-reliance. There are two parts to the thesis. The first part claims that principles—laws, obligations, social or political practices—that affect us bind us only when we can, in principle, consent to them. Specifically, the first part of the thesis of individualism stipulates that we must respect rational autonomy and that no person ought to be 'compelled' to accept principles that he or she cannot rationally identify with. Part two of the thesis states that society, or a scheme of cooperation, or the principles of social relationship respect the separateness of persons when they do not collapse a person's conception of the good with those of others.

Insofar as the activities of the state satisfy the two theses, its authority is acceptable and justified. There is a sense in which the two theses are related. That is, there is a sense in which a scheme that satisfies the first thesis would more than likely satisfy the second thesis. It is most likely that a scheme of cooperation that does not benefit the people or that places some people at a disadvantage at the expense of benefiting others would not command the willing acceptance of those that have been disadvantaged. Conversely, a scheme of cooperation that satisfies the demand of mutual advantage would most likely command the willing acceptance of those it benefits. The idea is that it is difficult to convince any person to accept a particular social scheme or to justify before an individual that a scheme of cooperation respects the equal moral status of persons or maximizes mutual

interests if that scheme places that individual at a disadvantage. How do classical social contract accounts argue for the two theses?

The arguments that these accounts provide for the theses follow this general outline. First, they specify a scheme that consists of political government, i.e. a state that realizes 'genuine liberties.' Next, they contrast this scheme with a state without political government, i.e. a state of nature and of 'unsecured liberties.' This state is described as a state marked by a general level of insecurity and insufficiency of resources; a state of universal noncooperation. Conditions and outcomes in this state are described as suboptimal and zero-sum, either because there are no guarantees for the benefits or liberties that the people exercise or the conditions for the full realization and expression of their freedom are lacking. Conditions are zero-sum when the gain or loss of X is exactly balanced by the loss or gain of Y. If the total gains of X (or Y) are added up, and the total losses of Y (or X) are subtracted, they will sum to zero.

In contrast to a state without political government, a scheme that consists of political government is regulated by the principles of social relationships. This state is marked by a general level of peace and security. Conditions and outcomes in this state are described as optimal and strictly non-zero-sum. Conditions and outcomes are strictly non-zero-sum when the aggregating gains and losses of X and Y are either less than or more than zero. That is, a situation is zero sum for X and Y when both gain or suffer together. In this state, the people are able to fully exercise their freedom and to pursue separately their business and interests. According to classical social contract accounts, because a social scheme that consists of political

government engenders outcomes that individually benefit everyone, the people will willingly consent to its authority.

1.1.2 Contractarianism and Political Government

Contractarian accounts (Hobbes and Locke) appeal to the need for the preservation and protection of property broadly construed, i.e. external property and property in our own bodies to justify the authority of political government. Hobbes bases his argument for the legitimacy of political government on our need for individual self-preservation, which according to him is the primary motivating factor behind the formation of political society or the commonwealth. Locke's argument for the legitimacy of political government is based on the need to protect the lives, liberty, and material possessions of those who lived within it. The reason humans would agree to form a political government and to accept its authority, Locke argues, is because its existence is necessary to the protection of their property.

1.1.2.1 Hobbes's Social Contract: Self-preservation, the Commonwealth, and the Leviathan

Hobbes is generally considered as one of the few truly great political thinkers. He is widely regarded as the founding father of modern political philosophy. His magnum opus, *Leviathan* rivals in significance the political writings of Plato, Aristotle, Locke, Rousseau and Rawls. For Hobbes, as it is for some of the other classical social contract theorists, the social compact is a framework for addressing the problems that exist in the state of nature. Hobbes takes the most basic problem

in the state of nature to be one of general insecurity; the other being insufficiency of resources. The sense of insecurity, according to him, is a fundamental reason why people seek to leave the state of nature. It is not clear what the extent of this insecurity is, for Hobbes, but it is obvious that he believes that it is sufficient to frustrate our self-interest, construed broadly as the desire for self-preservation.

For Hobbes, we are confronted with two fundamental choices: to continue in the state of nature or to join with others to form a commonwealth. The commonwealth is a state of optimality. In this state, liberties although restricted are guaranteed. By contrast, the state of nature is zero-sum and suboptimal. It is suboptimal because although we are free to exercise our liberties we are not certain of the state of our lives. Liberties are unrestrained, but are not guaranteed. As Hobbes sees it, the state of nature creates the problem of negative externalities in its most severe form.

The way Hobbes characterizes the state of nature makes this state the worst possible condition in which people can find themselves. Bear in mind that since this state exists whenever there exists a group of people who pursue their separate good and interests and recognize no group or body as having authority over them, the problem of negative externalities can hardly be resolved by scattered individual efforts. The pursuit of individual good and interests in the midst of unlimited equal power, according to Hobbes, leads to perpetual fear and a general level of insecurity, i.e. a state of war of all against all. This is worsened by the fact that moral concepts in the state have no objective grounding. As Hobbes puts it, whatever the object of any person's appetite or desire that is what that person calls

or judges to be good, and whatever the object of any person's hate and aversion that is what that person judges to be evil.[13]

If it is true, as Hobbes says, that, in the state of nature moral concepts are subjective, then the real nature of the conflict and war is a 'conflict of belief and judgment,' rather than one of acquisitions. Given the constant fear in the state of nature and thus uncertainty about self-preservation, each person makes the following three judgments: one, judgment about their position in the state of nature; two, judgment about the level and types of threats that they face; finally, judgment about what they need to do to minimize or remove the threats. In other words, individuals in the state of nature have to make some judgments concerning potential threats, whether or not their lives are at risk or have been endangered and what they should do to remedy the situation.

In the absence of a common authority to decide these matters, it is not surprising that Hobbes describes life in the state of nature as brutish, chaotic, and short. Given the dire condition in the state of nature, it is not unexpected that humans *qua* humans are moved to form a commonwealth. Hobbes' commonwealth is presided over by a common authority: an absolute sovereign (or Leviathan[14]) who is the final authority that decides among other things general matters of life and self-preservation. The sovereign ensures that the people honor the agreements

[13] Hobbes, *Leviathan* (1651), Richard Tuck (ed.), Cambridge, Cambridge University Press, 1996, ch.6, p39.

[14] The sovereign, for Hobbes, is not necessarily an individual. It may either be a single person or an assembly of people. Whether the sovereign is an individual or an assembly of people, it is, Hobbes believes, the representative of the people in the sense that it represents their "person," and it is the ultimate judge of what is conducive to human preservation. See ibid, ch. 18, pp.121-129 for Hobbes' discussion of the person and rights of sovereignty.

they entered into, in addition, of course, to punishing those who threaten the security of others.

Since Hobbes believes that the desire for self-preservation is sufficient to move people away from the state of nature to the commonwealth, it is important to recognize why he thinks a third-party mechanism is a necessary component of the whole process of the social compact. It is Hobbes' view that without a third-party mechanism whatever rights and liberties that the people have or are ascribed to them is null, if not empty and pointless. In the absence of guaranteed rights and liberties self-preservation would be in jeopardy and conditions would be as they were in the state of nature.

From the foregoing, it is clear what Hobbes' view is about costs and benefits. In natural human interactions, not only are marginal costs in excess of marginal benefits, but total costs are in excess of total benefits. But in the commonwealth, where there exists the mechanism to guarantee rights and liberties, total benefits are higher than total costs and for each person, marginal benefits are not less than marginal costs. Free from the need for constant vigilance against threats to their lives and property, the people can pursue their interests and happiness that benefit them and the commonwealth.

It is for this reason that Hobbes considers the contract which brings together otherwise morally unrelated individuals with separate interests and goals in the commonwealth the anchor for liberties and the basis for the effective preservation of life. The argument that self- preservation is only guaranteed in the commonwealth, that is the argument that the peaceful and stable environment that

the commonwealth makes available allows for the possibility of citizens to pursue freely their diverse interests and happiness is therefore an argument for the first and second theses. Thus, for Hobbes, because political government works to their benefits, citizens would willingly consent to its authority

1.1.2.2 Locke's Social Contract: Property and Civil Government

Locke, whose greatest philosophical contribution is his *An Essay Concerning Human Understanding* is widely regarded as one of the most influential Enlightenment philosophers in our modern world. His philosophy has had enormous influence on the development of epistemology and political philosophy. Political government, for him, as it is for Hobbes, is a product of the contract, which is justified on the basis of what it engenders—beneficial conditions for individuals. In Locke's social contract, the people agree to set up civil authority for the purpose of protecting the rights and liberties of everyone and to address property disputes that arise both in natural human interaction and in political society.

Property plays an important role in Locke's social contract theory. It is at the center of his argument for civil government because it is the protection of their property, broadly construed[15] that people seek when they choose to abandon the state of nature. So for Locke, as it is for Hobbes, the need for security, is a motivating force for people to institute civil authority. However, unlike Hobbes,' Locke's civil authority is not an absolute sovereign, but a non-extensive

[15] When Locke speaks of property he takes it to mean both external property such as land and property in our own bodies

government. The difference in their views on civil authority is partly explained by their different outlooks of human nature and the state of nature.

The state of nature, for Locke, is not a Hobbesian state of war of all against all. The state is pre-political but not pre-moral, that is it is not an amoral state. What does it mean to say that the state of nature is pre-political but not pre-moral? What it means is that it is not a state without morality and benevolence but a state that lacks a civil authority to punish people from transgressions against laws (of nature) and rights. Given that the social precedes the political, it follows that a group of people could have covenants or be governed by a system of agreements and still be in the state of nature. In this case, the group would constitute a society or a social community and yet not constitute a political society.[16]

According to Locke, the state of nature is relatively peaceful and people are free to pursue their diverse interests and goods. They are also free from social encumbrances and interference from others. However, to say that they are free to pursue their diverse interests does not mean that there are no limits on behavior. Again, the state of nature may be pre-political but it is not amoral or free from morality. Individuals in this state are constrained by a 'thin' sense of morality and by the 'law of nature' which, for Locke, is the same as God's immutable divine

[16] The view that a group of people may constitute a social community and yet not constitute a political society seems to suggest that for Locke the state of nature is a relational concept and it is only by *some agreement to join a legitimate political community with others* that takes people out of the state of nature. In § 14 of the *Second Treatise*, Locke says, "for it is not every compact that puts an end to the state of nature between men, but only this one of agreeing together mutually to enter into one community and make one body politic; other promises and compacts men may make one with another and yet still be in the state of nature."

law.[17] That is to say, humans have an innate knowledge of right and wrong and the law of nature and they are moved to a significant extent to act by this innate sense.

Although the state of nature is not the same as the state of war it can however collapse into a state of war, in particular a state of war over property disputes. In the absence of any civil authority to appeal to when there is a dispute over property individuals are allowed to employ any means to defend their own lives however, they see fit, including killing those who bring violence and force against them. Moreover, since people are the sole interpreter of the law of nature, there could be dispute as to whether the law of nature has been violated and the extent of such violations. In the absence of an impartial authority or judge to refer such matters to the dispute may remain unsettled and may cause contention and conflict.

In addition, when people take into their hands the punishment of alleged violators they could be biased toward their own interests, make the violations worse than they actually are and inflict excessive punishment. The institution of a civil authority prevents this by transferring all the interpretation, execution and punishment into a single, common and impartial authority. Again, as is with Hobbes as well as Locke, it is only in virtue of political government that rights and property are preserved. The cost of accepting the authority of political government is rewarded by the protection of property and rights, and for this reason the people, Locke argues, will voluntarily accept to be under it.

[17] In §6 of the *Two Treatises*, Locke says that 'the state of nature has a law of nature that governs it and this law obliges everyone to behave morally.'

1.1.3 Contractualism and Political Government

The authority and legitimacy of political government are defended by accounts of contractualism (Rousseau and Kant) on the ground that it makes available the conditions for the full and genuine expression of freedom. Rousseau bases his argument on the general will, which for him is the same as popular sovereignty. Citizens, he argues, express their full freedom when they create themselves anew in the general will. An individual expresses her full freedom when she joins with others in creating the laws that they are to obey. And when citizens come under the general will they put themselves under a collective body in lawmaking, i.e. they make laws for themselves.

Kant, on the other hand, grounds his justification of state authority and power on the need to respect the equal moral importance of persons as rational autonomous beings. All rational beings, he claims, have an 'innate right' to freedom. In addition, they have an obligation to enter into a civil condition governed by a social contract in order to realize and preserve this freedom. In the state of nature, rational beings are unable to realize and preserve their innate right to freedom because they are governed not by autonomous laws that they rationally give to themselves, but by desires and inclinations or some eternal conceptual relations that hold true independently of them. When citizens fail to govern themselves by laws that they give to themselves they treat themselves and others as mere means to the realization of some given empirical good and fail to respect their equal moral importance as rational autonomous beings.

1.1.3.1 Rousseau's Social Contract: Human Freedom and the General Will

Rousseau is one of the most influential philosophers of the Enlightenment. His social and political writings had enormous influence on the French Revolution and the development of contemporary social, political and educational thought. The social contract, which justifies political government, seeks, according to Rousseau, to remedy the social and moral ills that have been created by the 'progress' of civilization, i.e. the degenerative stage of society. The social contract transforms individuals with private or particular wills (*volunté particulière*), which aim at private interests to the general will (*volunté générale*) or the will of 'the whole community,' which aim at general interests or the common good. Rousseau announces the fundamental problem to which the social contract provides a solution at the outset of *The Social Contract*:

> To find a form of association that will defend and protect the person and goods of each associate with the full common force, and by means of which each, uniting with all, nevertheless obey only himself and remain as free as before.[18]

Why does Rousseau think the social contract creates for every person the most advantageous form of association? And what would it mean to have an association that protects the person and goods of each one, and yet preserves the freedom of that person and everyone? The answers to these questions can be found in Rousseau's idea of genuine or essential freedom. Genuine freedom, for Rousseau, is tied to the general will which is the same as popular sovereignty and which is

[18] Rousseau, *Of the Social Contract*, bk. 1, ch. 6, § 4.

always directed at the public good. Because the general will is always directed at the common good, it speaks infallibly to the benefit of the people.

The contract, Rousseau argues, describes the mechanism by which moral transformation or transition takes place—a transition of the *primitive individual* to *citoyen*. Specifically, a transition from *noble savage*, that is activity purely guided by self-interest or particular will to a political body, i.e. activity that expresses the individual's real or rational will. For the transformation to take place it is not enough, according to Rousseau, for the individual to set aside her egoism or private and particular will which works against the full realization of her freedom. In addition to setting aside her egoism, the individual must unite with others in the process that gives rise to the emergence of the general will which expresses her complete freedom. Rousseau speaks of the transformation in this way:

> As soon as this multitude is thus united in one body, one cannot injure one of the members without attacking the body, and still less can one injure the body without the members being affected. Thus duty and interest alike obligates the contracting parties to help one another, and the same men must strive to combine in this two-fold relation all the advantages attendant on it.[19]

There is an important reason why Rousseau thinks the transition from the *primitive individual* to *citoyen* is fundamental to genuine freedom. The transition, in his view, is the means by which humans reconcile their truly essential freedom with how they live together. Note that, for Rousseau, this transformation is both backward and forward looking. It is backward looking in the sense that it is

[19] *Social Contract,* bk. 1, ch. 7, § 4.

grounded on how humans are in the state of nature. It is forward looking in the sense that it speaks of what humans are now or will be in the future, i.e. a 'new person.'

The state of nature, for Rousseau is peaceful and free of vice.[20] In this state, humans are not amoral, neither virtuous nor vicious. They are *primitive individuals* who are isolated and peaceful. They are timid without any developed potentials to project and worry about the future. In addition, they have neither legal nor moral authority nor natural right to govern others. Any power that they exercise over others is established only by force or coercion. Rousseau encapsulates the free and innocent condition of humans in the state of nature which was subsequently supplanted by the corrupted condition in the degenerative stage of society in this famous statement, "Man is born free, and everywhere he is in chains. One man thinks himself the master of others, but remains more of a slave than they." What does Rousseau mean by this?

The progress of civilization, according to Rousseau, substitutes the free and innocent condition people enjoy in the state of nature for economic and social inequalities and subservience to others through dependence. To understand this, let us look at the distinction between *amour de soi*—a positive self-love—and *amour-propre*—a negative self-love. *Amour de soi* refers to the instinctive disposition of sentiment of self-preservation that human beings have in the state of nature. It is

[20] Just like most philosophers of his day, Rousseau understood the state of nature as a hypothetical device and looked to it as a normative guide. In the preface to the *Discourse on the Origin and Basis of Inequality Among Men,* he says, "The man who speaks of the 'State of Nature' speaks of a state, which no longer exists, which may never have existed, and which probably never will exists. It is a state of which we must, nevertheless, have an adequate idea in order to judge correctly our present condition.

the sentiment that is exclusively directed towards oneself as an absolute and valuable existence. Although acts of *amour de soi* are directed towards individual wellbeing, they are not malicious in the sense that they do not involve pursuing one's self-interest at the expense or detriment of others. On the other hand, *amour-propre* is artificial and generates vicious and competitive passions, and encourages individuals to compare themselves with others. It is a sentiment or disposition to be competitive and to measure one's activity by the activities of others. Acts of *amour-propre* are malicious in the sense that they are directed towards feelings of pride and competition.

What Rousseau seems to be saying is that in the state of nature humans were governed by *amour de soi*, but this is transformed into *amour-propre* by the 'progress of civilization.' In this state, humans have the disposition to frequently compete with others, while at the same time becoming increasingly dependent on them. The double pressure threatens both their survival and their freedom. Since the state of nature is neither feasible nor desirable in the light of where they are they have to come together to form an association that will defend and protect their goods and guarantee their freedom. Thus, by joining together into civil society through the social contract and by abandoning their claims to natural right, humans are able to both preserve themselves and to remain free. This is because submission to the authority of the general will as a whole guarantees that their wills are not subservient or subordinated to the wills of others and that they are not dependent on others. In addition, accepting the authority of the general will ensures that they obey only themselves because they are, collectively, the 'authors of the law.'

The purpose of the social contract, then according to Rousseau, is to reconcile our truly essential freedom with how we live together. Through agreement created with other free, rational and equal persons, the social contract reconciles our truly essential freedom with how we live together by subsuming our particular wills into the general will. This agreement is created through the collective renunciation of the individual rights that we enjoyed in the state of nature. We become *citoyens* when we abandon our natural rights and transfer them to the collective body. By transferring our natural rights to the collective body, a new 'person' is created. That is, we create ourselves anew as a single body, directed to the good and interests of all considered together. Thus, civil society in virtue of the social contract unites otherwise morally unrelated individuals by collapsing their good and interests under a collective body.

Rousseau contrasts the general will with the private or particular will and 'the will of all.' Private wills, which are ascribed paradigmatically to individuals in the state of nature, aim at private interests, while general wills, ascribed paradigmatically to the whole community aim at general interests, and collective preferences that aggregate preferences represent 'the will of all.' It is for this reason that Rousseau thinks that there can only be genuine authority 'when laws are made by individuals expressing their opinions not about particular will, but about what the general will is.'[21]

[21] Andrew Levine, *The General Will: Rousseau, Marx, Communism*, Cambridge, Cambridge University Press, 1993, p.18.

To illustrate what the general will is and how it contrasts with particular wills and the will of all, let us consider a firm of 100 employees.[22] Suppose the firm has a fixed sum of $1M available for workers' bonuses. Since particular wills are registered in private interests, it is to each employee's interests, i.e. particular will to get as much of this money as possible. We get the will of all when we add these particular wills. If we suppose that each employee wants the entire $1 million, the will of all will be a policy that distributes $100M. But $100M is not on offer for workers' bonuses. We recognize here that any claim put forward cannot be grounded on the pursuit of private will, which does not work to the interest of all. There is no $100M to be distributed and for each to insist on a policy of private will is to miss out on the $1M available for distribution. Given that each person is rational and equal with others and would need to justify his or her claim before others, each will put forward a claim of $10,000 for the $1M that is on offer. The result represents the general will, which is a policy that is equally in the interest of all the employees.

Whenever the people's opinion is not about the general will it is their private wills registered in their preferences that is at play, and this Rousseau claims is contrary to demands of reason and freedom. The general will tends to the public good and it decides on matters that are of public or general interests. Although the general will is not an aggregation of private individual wills, it is important that the opinion of each and everyone represent general interests and the collective body is preserved to the degree everyone's opinion reflects general interests. This explains

[22] The substance of this example is drawn from Jonathan Wolff's *An Introduction to Political Philosophy* (revised edition) Oxford, Oxford University Press, 2006, p.79.

why Rousseau believes that the laws enacted by the general will must be enforced. He writes, "whoever refuses to obey the general will shall be constrained [or compelled] to do so by the entire body: which means nothing other than that he shall be forced to be free."[23]

But doesn't this violate the second thesis—the thesis of individualism, i.e. the demand that no person ought to be *compelled* to accept practices that such a person cannot rationally identify with and that every person must consent to practices that affect him or her? How does Rousseau justify the claim that 'whoever refuses to obey the general will shall be compelled to do so by the entire body?' His explanation is that since an individual is only truly free when that individual identifies his or her interests with those of the general will, an individual's freedom is promoted and fully realized when he or she is forced to obey the general will. Being forced to obey the general will is not to diminish or remove one's freedom, for the authority of the general will is not a limit on freedom, but an expression of it. One might worry though if Rousseau is right to equate freedom with obedience, or more appropriately 'forced obedience.' X might have taken part in making a law, i.e. in the decision-making procedure and to this extent it seems justified to compel X to obey the law, but it is preposterous to conclude from this that such coercion makes X truly free.

It is clear from the foregoing that Rousseau has indeed created a social contract that fosters a community spirit, which on the face of it violates at least the second part of the thesis of individualism. For in transforming ourselves from *primitive individuals* to *citoyens* we give up our purely self-directed interests for

[23] *Social Contract,* bk. 1, cp. 7, § 8.

the interests of the whole community. However, in doing so our interests are wedded with the interests of others. The general will is a condition of 'total alienation of each associate with all his rights to the whole community.' Illustrative of such community spirit are soldiers or people dying for their countries in times of war and conflict; people giving themselves up to be tortured for the good of their societies; people willing to suffer variously to preserve the national security of their countries.

Rousseau's view about the general will thus raises a fundamental problem for his social contract theory. We can frame the most pressing problem that arises for his account of the social contact in terms of the apparent tension that exists between liberal individualism and communitarianism. Liberal individualism claims that individuals enter into society to further their interests, without taking the interests of society into consideration. Communitarianism accepts part of this claim. It accepts that individuals enter into society to promote their interests, but emphasizes that the interests of individuals need to be balanced with those of the community or society. Balancing individual interests with the interests of society may require an individual to sacrifice her interests for a higher cause, namely, that which promotes the common good. Rousseau's social contract is a species of communitarianism. On the one hand, Rousseau argues that the general will promotes individual interests, diversity and freedom. But at the same time, the general will also encourages a community spirit and the promotion of the common good. But promoting the wellbeing of the whole community can conflict with the

particular interests of individuals. How can the general will consistently promote the wellbeing of the whole and the interests of the individual?

Perhaps this tension can be dissolved by appealing to the distinction often made between two types of altruism, namely: 'collective altruism' and 'de facto altruism.' An act is collectively altruistic when it aims towards everyone, that is, when the act benefits every person including those that perform it. On the other hand, an act is de facto altruistic when it does not aim towards everyone, that is, when others and not those that perform it are not benefitted by it. Indeed, given the collectivism of the general will, one might say, it is at bottom altruistic. In which case it may be said that Rousseau's point seems to be that *ipso facto* the general will is concerned with equating individual's interest with that of the collective it gives rise to 'collective altruism' rather than 'de facto altruism.' In collective altruism, no one seeks to take advantage of others, but seeks to benefit others as well as themselves. But how is this possible? How can it be said that one benefits others at the very time one seeks one's own interests? Specifically, how is it possible for an individual to seek her interests and yet be collectively altruistic, especially when her interests conflict with those of the collective?

Rousseau's answer is that the individual's interests, which are explained by his or her rational and essential freedom are realized by the general will. The general will is not the simple collection of individual wills, but the expression of an individual's true and rational will in the sense that it furthers that individual's interests, namely, it realizes that individual's freedom. To put this in a different way, an individual's interests are satisfied when that individual creates herself

'anew' in a common body, which realizes her essential freedom. Given that what the individual seeks is the satisfaction of her individual interests, and given that the general will is directed to general interests which includes the interests of that individual, the individual satisfies her interests then when she creates herself 'anew' in a common body. Thus, when she acts in accordance with the general will the individual benefits others in the very same moment she benefits herself.

If you put forward a claim of $10,000 for the $1 million that is on offer, you benefit yourself: rather than nothing, you have $10,000. However, in benefiting yourself you benefit others as well: they individually have $10,000 and not nothing. If one were to benefit others without benefiting oneself, one is a de facto altruist, but if one were to benefit others the very same moment one benefits oneself, one is a collective altruist. Thus, by transcending themselves collective altruists learn the true meaning of duty and justice, where the voice of duty, according to Rousseau, "succeeds physical impulsion and right succeeds appetite, and those, who until then had looked only to themselves, see themselves forced to act on other principles, and to consult their reason before listening to their inclinations."[24]

If the true meaning of freedom is self-transcending, then Rousseau is right to claim that self-transcending leads to self-mastery or self-control. Generally, a person that has self-control is one that is not easily sidetracked by side attractions. Self-control requires not giving in to temptations. The sense in which we have self-control, for Rousseau, is not different from the common view of self-control, even though he applies it narrowly to the general will, namely, when we transcend

[24] Ibid, bk. 1, ch. 8, § 1.

things opposed to the general will: mundane interests, inclination and appetite. We have self-control when we act in accordance with duty and justice and we act in accordance with duty and justice when we subsume our private wills under the general will.

Note that Rousseau's answer does not dissolve the tension that seems to exist between liberal individualism and communitarianism. Particularly, it does not remove the worry that his social contract lumps together the various conceptions of the good of citizens and treats a group of many as if it were a single person. Rousseau's answer only explains how one's interests are conceptually tied with the interests of others. That one's interests are conceptually linked to common interests does not remove the fact that general interests can still conflict with one's particular interests. It is a fact that we belong to different religions. We have different histories and come from diverse racial, cultural, social, and ethnic backgrounds. Also, we have different moral, economic, and philosophical ideas. Since these affect our values, it is likely that we would value different things even though we all have similar basic needs. Some may value economic progress while others value the protection of the natural environment. Also, some may value a minimal government while others value a government that is involved in social programs or programs of wealth redistribution. Because we may value different things, there are bound to be conflict of interests. Thus, on many important issues it is quite unlikely that there could be any policy that is beneficial to all or that is equally in the interests of everyone.

The morals by decision-value agreement account I shall be defending in this essay provides a middle way between liberal individualism and communitarianism in general and between the strict individualism of Gauthier's morals by constrained maximization agreement and the extreme communitarianism of Rousseau's social contract theory. Mb(CM)A grounds morality on rational self-interest, which is identified with the maximization of expected utility. Morals by decision-value agreement account identifies rationality with the maximization of decision-value. As a species of communitarianism, Rousseau's social contract is extremely collectivistic because it identifies individual interests with collective interests. Gauthier's moral contractarianism is a species of liberal individualism. It is highly individualistic because it identifies individual interests with the maximization of that individual's expected utility. Mb(DV)A navigates between these two views. Mb(DV)A identifies individual interests not with collective interests or the maximization of expected utility, but with the maximization of decision-value. Under Mb(DV)A, our utility profile includes our preferences and aversion for practices and actions, in addition to the expected utility of those practices and actions, such that what we maximize when we act is not the expected utility of those practices and actions, but the value of those practices and actions. For Mb(DV)A we are individuals *disposed-communally*, that is we are disposed *individually* and *communally*. Specifically, Mb(DV)A claims that by maximizing decision-value we maximize both our interests and the interests of others.

1.1.3.2 Kant's Social Contract: Rational Freedom and the Commonwealth

Kant is widely regarded as one of the most influential thinkers throughout the history of Western philosophy. His contributions to metaphysics, epistemology, aesthetics, moral and political philosophy are enormous, and have profoundly influenced almost every philosophical movement after him. The aim of his social and political writings is to champion the Enlightenment in eighteen-century Europe in general and the idea of freedom in particular. For Kant, the social contract is a rational justification of state authority. As we will see with Rawls later, Kant thinks that the social contract seeks to discover fundamental moral principles that will govern political society. These are principles that match what Kant refers to as our 'innate right' to freedom. By 'innate right' to freedom, Kant means "independence from being constrained by another's choice, insofar as it can coexist with the freedom of every other in accordance with a universal law."[25]

Since persons, according to Kant, are separate entities, they must be respected as ends in themselves. Respecting them as ends require that their conception of the good or happiness be treated independently of the conception of the good or happiness of others. This explains why he rejects any other basis for political government or why he argues against grounding state power on the welfare of citizens. The government, he argues, cannot justifiably impose any particular conception of the good upon its citizens. To do so would be for it to treat citizens not as fully autonomous persons but as children. To treat them as such is to suppose that they are unable to comprehend what is truly beneficial or harmful to themselves. Kant's claim that the government cannot impose any particular

[25] Kant, *Metaphysics of Morals*, 6:237

conception of the good upon its citizens must be understood in the light of the more general claim he makes in moral philosophy. This is the claim that the moral law or the categorical imperative is not heteronomous, by this he means that the moral law must be autonomously given and cannot be based upon happiness or any other empirical good.

We can understand the sense in which state authority for Kant is justified if we look at what he thinks justifies the principles. Every rational person, Kant claims, must appropriate as part of his or her rational nature the principles that arise from the social contract. By this he means that every person must consent to the principles that are to govern political society. The principles are not heteronomous in the sense that they are imposed neither by our psyches nor by some eternal conceptual relations that hold true independently of us. To the extent that the principles are not imposed by some eternal conceptual relations but arise autonomously from us they are congenial to our conception of the good. Because they are not independent of our conception of the good, they respect our equal moral status as autonomous persons.

The political government that the contract creates, just like the principles must respect our equal moral status as autonomous persons. The 'original contract' Kant says is an 'idea of reason' that compels everyone to act in ways that coexist with everyone's freedom, in accordance with a universal law. To this extent the sovereign is obligated to "give his laws in such a way that they could have arisen from the united will of a whole people and to regard each subject, insofar as he

wants to be a citizen, as if he has joined in voting for such a will."[26] A commonwealth that respect our equal moral status as autonomous persons distributes advantages and burdens equally. Everyone's freedom is constrained by everyone's freedom, and no one is placed in a position to impose their own freedom or conception of the good on others.

The commonwealth, Kant says, lifts humans from the degraded condition of the state of nature. Humans in this state are debased and constantly engaged in a war of all against all. It is a barbaric and brutish state that gives free rein to the "freedom of folly,"[27] for a "state of peace among men living together is not the same as the state of nature, which is rather a state of war" and anarchy.[28] For even if conditions in the state of nature do not involve active hostilities and conflict it involves, according to Kant, "a constant threat of their breaking out."[29] Given that the state of nature hinders our exercise of genuine freedom, humans have a duty to enter into a civil condition governed by a social contract in order to realize and preserve their innate right to freedom. And insofar as the institution of the commonwealth or the state creates the very condition for the possibility of realizing and preserving this freedom, it respects citizens as autonomous persons and ends in themselves, and to the extent it does this its authority is justified.

Since the social contract reflects reason, each citizen as a rational being already contains the basis for rational agreement to the authority of political

---

[26] Kant, "On the Common Saying: That May be Correct in Theory, but it of No Use in Practice," 8:297.
[27] *Perpetual Peace: A Philosophical Sketch*, 2nd §, p.103, ***Kant's Political Writings***, H.S. Reiss (ed.) Cambridge, Cambridge University Press, 1991.
[28] *Perpetual Peace,* 2nd §, p.98.
[29] Ibid, p.98.

government and hence may be forced, according to Kant, into the commonwealth against his or her consent. Given that the citizens recognize the social contract as embodying their 'ideal self' or rational being since it creates the condition for them to realize and preserve their innate right to freedom, they would subject themselves—voluntarily or coercively—to rational agreement to the authority of political government. Unlike Hobbes, who centers his argument for the justification of the authority of political government on the individual benefit for each person, Kant bases his argument on the innate right to freedom of everyone in general.

1.2 Moral Contractarianism and Utility-Seeking Agents

Gauthier's moral contractarianism has its origin in Hobbes and it contrast with Rawls' contractualism which has its root in Kant. I said in the introduction to this chapter that contractarianism is different from contractualism in some relevant sense. Whereas contractarianism is based on mutual self-interests, contractualism is grounded on the equal moral status of persons. Under contractarianism, what I aim for is the maximization of my interests in a bargain with others, who also seek to maximize their interests. However, under contractualism, what I seek is the pursuit of my interests in a way that I can rationally justify to others who have their own interests to pursue. My discussion of classical social contract theory demonstrated this difference, but it is with the social contract accounts that Gauthier and Rawls defend that the difference is most emphasized.

Rawls and Gauthier are respectively the most well know exponents of contractualism and contractarianism. They both apply the resources of the science

of rational theory that emerged in the 20th century to demonstrate how the principles of social relationships that arise from the contract contours individual behavior and society. Gauthier applies the resources of rational choice theory to the choice of the principles as a rational response to the problems raised for utility-seeking agents. Rawls, on the other hand, whose contractualism I shall examine in chapter two, applies the resources of rational choice theory to the choice of the principles as a way of demonstrating what the basic structure of society will be if the principles are chosen from a position of moral equality. This is not to say that self-interest does not play a role in Rawls' contractualism. He employs self-interest behind the veil of ignorance to represent a commitment to justice, construed as fairness to all. Because we know nothing about our personal characteristics or social and economic circumstances and because we know that we could end up being anyone once the veil is lifted, we must have concern for all.

In the various accounts of the social contract that Gauthier and Rawls defend, mutual advantage plays a significant role. Society, they claim, is 'a cooperative venture for mutual advantage.' Although mutual advantage plays a significant role in their accounts of the social contract, they differ regarding the type of social arrangement that satisfies it. For Rawls, mutual advantage is satisfied when we demonstrate what the most just and feasible arrangement of basic social institutions that realize the core democratic values of liberty and equality of all citizens would be. To demonstrate this is to delineate the scope of justice, construed as fairness for all. In Rawls' theory, agreements do not create morality even though they may generate moral and political obligations. And constraints

whether moral or rational do not arise from preferences but are either extended or transformed by the contract situation. For Gauthier, mutual advantage is satisfied in an essentially just society. An essentially just society is one that is both constituted by moral principles that speak to our considered preferences and that takes seriously the role that initial factor endowment plays in determining benefits and desert. An essentially just society takes seriously the role that factor endowment plays in determining what we receive when it proscribes activities that place some at a disadvantage for the benefits of others.

Gauthier calls his social contract theory *Morals by Agreement* to demonstrate that moral constraints are circumscribed by agreement of rational persons for the purpose of advancing their rational self-interests. Morals by constrained maximization agreement jettisons some of the assumptions and theoretical underpinning Rawls employs in JaF.[30] One of them is the capacity for a sense of justice, by which Rawls means that we can be relied upon to comply with principles that we choose in the original position; another is the assumption that we choose the principles behind a veil of ignorance. According to Mb(CM)A, moral constraints are the outcome of a bargaining process by utility-seeking agents, by this Gauthier means that constraints arise from the preferences of bargainers who seek to maximize expected utility.

To take constraints as arising from our considered preferences is to ground moral principles on our rational self-interest. And by arguing that constraints maximize our rational self-interest explained by our preferences, Mb(CM)A argues

[30] I will be using JaF interchangeably with ToJ to refer to Rawls' contractualism. For the most part, I will be using ToJ to refer to Rawls' book, A *Theory of Justice*, and JaF to refer to his contractualism.

that rationality requires compliance with agreements or moral principles. For if the contract provides the framework for agreements to emerge and if the agreements encourage cooperation and thus advance our rational self-interest, then rationality requires that we comply with them.

Moral contractarianism measures the extent of the constraints on utility-seeking and maximization by the benefit each person receives from doing so. The benefit to each person is measured by the total benefits available through cooperation and each person's contribution to it. If we take the total benefits (cooperative surplus) available as a single transferable good, then each person's contribution is identified partly with the initial factor endowment that the person brings to the bargaining table.

By arguing that rationality requires compliance with agreements, Gauthier's contractarian approach demonstrates that it takes seriously the rationality of the rational skeptic. Since the rational skeptic is motivated to break agreements when it advances his self-interest, his behavior, threatens the foundation of morality. As noted in the introduction, the problem posed for morality by the rational skeptic, which is the same with the general problem of rational compliance is essentially the Prisoner's Dilemma.

Traditionally, the Prisoner's Dilemma (PD) arises where the equilibrium non-cooperative outcome diverges from the optimal cooperative outcome. The PD has two persons awaiting trial presented with the following offers or options. If one prisoner confesses and the other does not, the one that confesses walks away free (0 years in jail) and the second receives some years in jail (say, 10 years of prison

sentence). If both confess, each receives some years in jail (say, 5 years). If both refuse to confess, each receives some jail time (say, 2 years of prison sentence). In any form of the problem or game, the choice is between cooperation (not confessing) and noncooperation (ratting or confessing). The problem can be represented as follows:

***Figure 1.2a*: The Prisoner Dilemma with Algebraic Variables**

		Prisoner 1	
		Don't confess	Confess
Prisoner 2	Don't confess (Cooperate)	n1, n2	m1, p2
	Confess (Rat)	p1, m2	o1, o2

Note: m1 > n1 > o1 > p1 and m2 > n2 > o2 > p2

***Figure 1.2b*: The Prisoner Dilemma with 'Magic' or Real Numbers Showing Years**

		Prisoner 1	
		Don't confess	Confess
Prisoner 2	Don't confess (Cooperate)	2, 2	10, 0
	Confess (Rat)	0, 10	5, 5

In classic form of the problem, cooperating is always strictly dominated by defecting, so that the only possible equilibrium for the game is for each player or actor to defect. That is, the dominant action is for each person to rat no matter what the other person does, even though that leads to a suboptimal outcome for each person. In iterated versions of the game, when the game is played repeatedly and the actors know in advance the number of steps, there exists the opportunity for each actor to punish the other for previous non-cooperative behavior. However, rational choice theory says the choice of each actor should not be affected by the repeated nature of the game. On this view, no matter how many times the game is played, each actor should defect repeatedly. The only time it is rational, according to rational choice theory, for actors not to defect is when the game is iterated infinitely, that is, when it is played endlessly or for a random number of times. Cooperation can be in equilibrium when the game is iterated infinitely because the motivation to defect can be overcome by the perpetual threat of punishment.

In informal usage, the PD may be applied to situations not strictly satisfying the formal criteria of both the classic or iterative forms of the game. An example of a situation not strictly meeting the formal standards of both the classic or iterative forms of the game would be one in which two entities (individuals or societies) could benefit from cooperating or suffer from not cooperating, but find it difficult to coordinate their activities to achieve cooperation. In its formal usage, the focus is on the problem that rational defection poses for cooperation, whereas in its informal usage, the focus is on the problem posed for cooperation by the difficulty of coordinating activities.

If the rational skeptic can "bait' others into honoring their part of a bargain or to produce the benefits of cooperation while avoiding the costs at the same time, he would do so. By reaping the gains of cooperation while avoiding simultaneously the costs that generate them, the rational skeptic prefers a situation where the other player chooses not to confess while he rats. If the other prisoner agrees to cooperate by not confessing, and he chooses to rat, his choice would greatly benefit him (earn him freedom — 0 years in jail), and the other prisoner 10 years of jail time. By his behavior, the rational skeptic thus demonstrates that he does not accept the rationality of mutual morality.

In MbA and elsewhere, Gauthier argues for compliance with moral principles by appealing to rationality itself. He develops a conception of rationality that he says is 'capable of withstanding critical examination,' and then identifies the constraints imposed by that conception of rationality with moral principles or morality. The result is Mb(CM)A, a moral theory that Gauthier claims is not only 'compatible with the conception of rationality' he has developed, but a moral theory that he argues 'offers the only plausible resolution to the foundational crisis facing morality in our modern world.'

Moral contractarianism, Gauthier argues, resolves the PD. It does this with the idea of constrained maximization—the strategy to comply with mutually advantageous moral constraints. A person is a constrained maximizer if that person is disposed to comply with mutually advantageous moral constraints, provided the person expects similar compliance from others who are similarly disposed. On this view of rational morality, what motivates us to honor agreements is that they

present themselves to us as necessary instruments by which self-interest is advanced.

How is it in one's interest to comply with agreement one voluntarily entered into, even though on a particular occasion breaking them may seem more beneficial? Stated differently, "what rational motivation do I have for not stealing your property, when I can get away with it? And what rational motivation do you have for not going back on your word, when it seems advantageous for you to do so?" A helpful PD illustration is this: suppose we both enter into a contract whereby we agree to help each other on days that we have chosen to see our favorite band play. The agreement requires each of us to babysit for the other. You will babysit for me on the weekend my band comes to town, and I will babysit for you the following weekend when your band is in town. But I am better off not babysitting for you after you have babysat for me. I could, for example, pick up extra work hours rather than babysit for you and therefore increase my gains. Since I am better off reaping the gains of cooperation while avoiding, at the same time, the costs required to produce them, what stops me from refusing to babysit for you next weekend after you have babysat for me this weekend?

One natural response is to say surely you would not babysit for me because you would not trust that I would babysit for you after you have babysat for me. You will know this and I will know this. In addition, both of us are aware that were I to babysit for you first, you will not return the gesture, therefore, you will not babysit for me and I will not babysit for you. The bottom line is that since both of us know that I would not babysit for you after you have babysat for me you will

not babysit for me in the first place. And since we would not babysit for each other, there would be no cooperation. Here we have the PD as figure 1.2b illustrates.

Figure 1.2c: **The Prisoner Dilemma with Matrix Showing Utilities**

		Me
	Babysit	Don't Babysit
Babysit (cooperation)	25, 25	0, 35
You		
Don't Babysit (non-cooperation)	35, 0	15, 15

Let us take the matrix entries in the above graph—which I first presented in the introduction—as utilities. The higher the utility the more gain each of us receives. If we suppose that we are individually rational, then it is better for us to cooperate, i.e. babysit for each other. For if we cooperate we each have 25 utilities compared to 15 utilities when we fail to cooperate. To cooperate is to comply with agreement and to comply with agreement is to constrain one's behavior by moral principles. To constrain one's behavior by moral principles is to promote a scheme that satisfies mutual advantage. A scheme of cooperation satisfies mutual advantage just in case such scheme works to the benefit of everyone. In the absence of mutual advantage, agreements are empty and in the absence of binding agreements there is no cooperation. Since noncooperation provides us individually suboptimal outcomes (15 utilities each), we have to find a way to both prevent you

and I from breaking the agreement we have entered into and to coordinate our activities to achieve cooperation. The sub-theory of constrained maximization is meant to explain how this is possible.

How does the idea of constrained maximization solve the problem of compliance? It solves the problem, according to Gauthier, by showing that the dispositions it is rational for us to form are those that maximize expected utility. It is rational for you to choose dispositions that match and corresponds to the dispositions that I have chosen, and it is rational for me to choose dispositions that match and corresponds to the dispositions that you have chosen. If we both of us benefit from cooperation, then it will be rational that the dispositions we choose are those that would enable us cooperate or that favor cooperation.

On this view of dispositions, your reason for complying with moral principles is explained in terms of the kind of dispositions that would be rational for you to form, if you expect to do better forming such dispositions. The reason you choose to babysit for me is that you expect to do better by keeping your commitment than you would have done if you have not made the commitment to babysit for me. The case then is simple: you will babysit for me because you know I would babysit for you. If you know I would not babysit for you, you would not babysit for me. But since I gain from you babysitting for me and since you gain from me babysitting for you, I would form the sorts of dispositions that would make me babysit for you and you would form the sorts of dispositions that would make you babysit for me. The dispositions enhance our willingness to constrain our

utility-seeking behavior and make it possible for us to coordinate our activities to achieve cooperation.

The constraints on utility-seeking behavior speak to our reason because they maximize expected utility. Because the constraints promote a scheme of cooperation that satisfies mutual advantage, Gauthier claims that they satisfy the demand of equality and fairness, which results when, for each person, marginal benefits are not less than marginal costs. The moral reasons for accepting constraints, according to moral contractarianism, lie in the maximization of our preferences explained by expected utility. To the extent that Mb(CM)A identifies constraints with self-interest and self-interest with the maximization of expected utility (EU) it requires that we do not extend our agreement more widely than benefits us.

The point about rationality forbidding us from extending our agreement more widely than benefits us seems intuitively compelling. If the condition for accepting constraints on our behavior is that they engender and promote a scheme of cooperation that satisfies mutual advantage, namely, they advance our rational self-interest, then it would seem rational and moral as well that we refuse any scheme of cooperation that does not benefit us. And were we already wedded into a scheme of cooperation or into a relationship with individuals who are now unable to benefit us, rationality tells us to break away from this scheme or relationship. We act rationally and morally when we exclude from the contract or any scheme of cooperation those unable to contribute to the cooperative surplus of that scheme. Simply put, because rationality or the rational morality of Mb(CM)A requires that

agents maximize EU it is not in the interest of better-off agents to interact with less well-off agents. This point about rationality requiring us to maximize expected utility and about excluding some people from agreement leads us directly to 'the problem of secession.'

1.3 Outline of the Problem of Secession and a Multi-tracked Framework for Solutions

A general test of the application for Mb(CM)A is 'the problem of secession': the problem concerning what should be done to previously endowed, productive, or better-off members of society who have for some reason become unproductive. Mb(CM)A's solution to this problem is demarcated by its conception of rationality. Mb(CM)A takes mutual advantage to be a cardinal feature of any scheme of cooperation. It defines mutual advantage as well as rationality in terms of the maximization of expected utility. This conception of rationality and its view of mutual advantage inform and define its solution in the test of application.

A scheme of cooperation that satisfies mutual advantage, according to Mb(CM)A, is one that is essentially just and an essentially just society is one in which people do not extend their agreement more widely than benefits them. Given that a scheme of cooperation involving productive and unproductive members fails to maximize the EU of productive members they would be acting rationally if they secede and form a society among themselves. In short, for Mb(CM)A, when EU is stacked too high against cooperation, it is not rational or in an individual's interest to cooperate.

The point is that if an entity (an individual or a group of people) that is better-off gets fewer benefits from cooperating with another entity that is less well-off it would be rational for the former not to honor contractual obligations or to support the latter since doing so works against the former's rational self-interest. Thus in situations of asymmetrical contributions or diminished or almost extinguished EU productive members are justified in excluding less well-off or unproductive members from cooperation. A scheme of cooperation consisting of better-off or productive members and less well-off or unproductive members is not mutually beneficial because it does not maximize the expected utility of the former. Since cooperation under this scheme or arrangement does not benefit better-off members it is rational and in their interest to secede from it and form a society of productive members.

Mb(CM)A's solution to the problem of secession is a single-tracked silver bullet solution. A single-tracked silver bullet solution tracks only a single 'value' or reason, and in the case of Mb(CM)A it tracks only EU reasons. We might contrast a single-tracked silver bullet solution with a multi-tracked framework for solutions. A multi-tracked framework for solutions tracks all moral reasons or values that play a significant role in an agent's decision or reasoning process. It tracks these reasons or values by factoring them into what counts as rationality and reasons for acting.

An account offers a silver-bullet solution to a problem if based on its single-tracked 'principle,' or 'value,' or reason it claims that it is rational for an agent to act one way or another in situations where acting one way or another is not

definitive or clear-cut. By contrast, an account offers a multi-tracked framework for solutions when on the basis of its value framework it claims that it is rational for an agent to switch between strategies or solutions, i.e. attach different weights to the available strategies or acts in situations where acting one way or another is not definitive or clear-cut.

The problem of secession may be described as one such situation, where acting one way or another is not definitive or clear-cut. The situation is not definitive or clear-cut because in addition to the possible outcome or expected utility of the acts of secession and non-secession, there are other utilities that are available to agents. These utilities attach to the acts themselves. Utilities attach to acts when the acts have certain values or when agents have moral reasons to choose the acts independent of their possible outcomes. Because what an act means for an agent is crucial to the choice of that agent, i.e. the act he or she chooses, the problem of secession is dissolvable by specifying the weights or values that an agent assigns to the acts of secession and non-secession and how those weights or values determine whether or not that agent cooperates with others.

Mb(CM)A's single-tracked solution may acknowledge these values or reasons but it does factor them into rationality. It does not explain rationality in terms of an agent's considered preference and aversion for the acts that are available. Rather, it explains rationality in terms of the maximization of expected utility. Rationality for it equals EU maximization. It is for this reason that it breaks down or fails when applied to the problem of secession. The breakdown of

Gauthier's brand of contractarianism in the test of application raises the following two related issues.

The first issue is that of 'naivety.' Mb(CM)A is infected and fixated with the idea—which runs through rational choice theory—that the maximization of expected utility is constitutive of rationality. Indeed, expected utility is a constituent of rationality. But if all we see when we look into reason or rationality is the maximization of expected utility, then we have certainly failed to look properly. If all we want about morality is put there by reason why should we think that reason or rationality equals the maximization of expected utility? Equating rationality with the maximization of expected utility is to have a stripped down notion of rationality and with it a naïve, narrow and misleading conception of practical rationality.

The second issue, which follows from the first, concerns Mb(CM)A's aim of deriving morals from reason. If it is the case that the account of rationality that Mb(CM)A offers is naïve, narrow, and misleading, then its task of deriving morality from rationality by identifying morality with rational constraints explained by the maximization of expected utility is problematic. Specifically, if what we see when we look into reason is the maximization of expected utility, then Gauthier's artful and sophisticated goal of identifying rationality with morality is in jeopardy, unless, of course, we revise such a narrow and naïve conception of rationality.

That is what I will be doing in this thesis. I shall be proposing a replacement of Mb(CM)A's conception of rationality with an account of rationality

that incorporates other moral reasons, i.e. a morals by decision-value agreement's account of rationality that takes rationality to be the sum of expected utility and the utilities or values that attach to acts. I identify a multi-tracked framework for solutions with morals by decision-value agreement, which I identify with a decision-value/symbolic utility account of practical rationality—the sort defended by Robert Nozick in the *Nature of Rationality*.[31] Decision-value/symbolic utility or morals by decision-value agreement (Mb(DV)A) explains the different moral reasons or values that play a role in the choice of agents not just in the PD but in choice contexts in general.[32] Because Mb(DV)A factors in the value or the meaning of secession and non-secession acts, and claims that rationality requires that agents maximize DV, it and not Mb(CM)A is able to dissolve the problem of secession.

[31] Nozick, *The Nature of Rationality*, Princeton, New Jersey, Princeton University Press, 1995.

[32] I shall be using DV's incorporation of SU in two ways. First, as Mb(DV)A, which would refer to DV factoring SU and EU in strategic or cooperative situations. Second, as decision-value/symbolic utility (DV/SU), which refers to DV factoring SU and EU in parametric or non-cooperative situations, and sometimes just as an account that factors EU and SU reasons in an agent's decision process. There is no theoretical significance and value to my using Mb(DV)A and DV/SU, other than that they serve as shorthand forms of a desire-based value-sensitive account, or as ways of differentiating between two choice situations.

Chapter Two

John Rawls' Social Contract Theory—*Justice as Fairness*

Introduction

I will examine in this chapter Rawls' contractualism. *Justice as Fairness* is the phrase used by Rawls to refer to his distinctive theory of justice. It is also the title of an essay on justice he wrote in 1958. JaF consists of two principles of justice: the liberty principle and the social and economic principle. Rawls is a 'contractualist' because he employs the device of the hypothetical contract to tease out the principles of justice that will structure the basic institutions of society and realize the core democratic values of liberty and equality of all citizens. He argues that the two principles would be chosen by representative parties (rational persons) placed in a condition of moral equality, in the original position.

I consider my discussion of Rawls' social contract theory important to my examination of Gauthier's moral contractarianism. JaF, I believe, serves as a fruitful background to Gauthier's brand of contractarianism because the latter draws on some of Rawls' ideas; including those of classical social contract theorists. Gauthier follows Rawls in employing the resources of the science of rational choice to demonstrate how fully rational agents come to agree on principles that are just and fair to everyone, notwithstanding the fact that the principles have different targets for them—to structure individual behavior, for Gauthier, and society, for Rawls. Gauthier pursues a different and more radical deduction project—strictly deducing morality from rationality—a deduction project that is implicit in Rawls but smeared by question-begging assumptions. The

assumptions that Rawls brings in to circumscribe the range of principles considered in the original position shows that the choice of principles for him would not proceed from rationality alone, but rather would be goaded by moral consideration, specifically by prior moral beliefs. Furthermore, Gauthier draws on Rawls' idea of society as a cooperative venture for mutual advantage, that is, the idea that individuals agree to cooperate or to be part of society because of the benefits that cooperation or society provides them. However, they both differ as to the type of society that satisfies mutual advantage. This difference is partly due to the different reasoning they employ in the pre-contractual stage. Whereas the reasoning in the original position for Rawls boils down to a single individual, the reasoning in the pre-bargain stage for Gauthier boils down to multi-players.

2.1 The Aim of Moral Philosophy or Theory[33]

Rawls agrees with Kant on the aim of moral philosophy or theory. Whereas, Kant contends that the aim is to seek out the fundamental or foundational principle of a metaphysics of morals, Rawls argues that it is to seek out the principles of justice. The purpose of moral philosophy, according to Rawls, is not to discover some absolute truths, which are 'out there.' Moral philosophy is not like physics that seeks to find out some kind of universal truths about our universe. Rather, moral philosophy is like grammar. The study of grammar does not seek to find some universal truths about language, but seeks to understand, analyze and clarify the way in which we use our native language. The same, according to Rawls, is true of

[33] Rawls, ToJ, pp. 40-46.

moral philosophy. The aim of moral philosophy is to understand, analyze and refine our moral beliefs and sentiments.

While Kant pursues his project by analyzing and elucidating commonsense ideas about morality, Rawls pursues his by appealing to our considered moral judgments. For Rawls, the method of refining our moral beliefs and sentiments is 'reflective equilibrium.' We start with our moral beliefs and convictions, not as they ought to be, but as they are. As humans, there are certain things that we can say we strongly believe. In Rawls's case, one of the most basic moral sentiments about society and justice is the belief and conviction that slavery and racism are wrong. Now, it is important to point out that the idea is not so much to try to argue or 'prove' that slavery and racism are wrong. The idea is to show that, for us, slavery and racism *are* wrong, and what we seek are principles of justice that incorporate these moral beliefs and sentiments. The method of reflective equilibrium is a process that goes back and forth between our strongly held moral beliefs or judgments and the principles of justice. It is an intuitive way of providing theoretical justification for our most strongly held moral beliefs or judgments, and of refining and redefining those judgments as is necessary.

Rawls does not employ the method of reflective equilibrium as a logical and scientific enterprise, rather he employs it as a method by which we find out what our strongest moral beliefs are, then try to construct abstract principles which are compatible with and justify them, and then refine those beliefs on the basis of those principles. If our moral judgments are in a state of balance or harmony with our overall principles, then these principles are in equilibrium. In view of the role

of the method of reflective equilibrium, the main thrust of Rawls' moral argument for the principles of justice, is to demonstrate that they are in reflective equilibrium with our 'considered moral judgments' about justice, or at least, that they are closer to being in a state of balance or coherence with those judgments than the main alternative theories of justice, utilitarianism and perfectionism.

In addition to the method of reflective equilibrium, Rawls employs a number of other assumptions in pursuing his project. These assumptions are central to the argument that the principles are those that would be chosen by representative parties in the original position. These assumptions are the 'veil of ignorance,' a device for screening out certain characteristics and information from the contractors in the original position; the maximin rule, the principle of rational choice that the contractors employ in choosing the principles of justice; the capacity for a sense of justice, the assumption that contractors can be relied upon to comply with the principles that they choose in the original position.[34]

Why would the liberty principle and the social and economic principle be chosen and not some other principles, say, some utilitarian principle of maximizing

[34] There is a sense in which the moral power of the capacity for a sense of justice relates to Rawls' point about ideal theory. For Rawls, ideal theory is lexically prior to non-ideal theory in the sub-domain of the political. He claims that completing the former first yields a systematic understanding of how to reform and improve our non-ideal world; ideal theory fixes a vision of what is the best that can be hoped for. Ideal theory makes two kinds of idealizing assumptions about its subject matter. First, it assumes that all rational actors are generally agreeable to comply with whatever principles that are chosen. This matches with the assumption that rational actors possess the capacity for a sense of justice. Ideal theory thus idealizes away the possibility of law breaking or of there being malcontents. Second, ideal theory assumes reasonably favorable social conditions, wherein citizens abide by principles of political cooperation. That is to say, ideal theory assumes that citizens are not driven by hunger or some other social and economic ills, for example, such that their capacity for moral reasoning is overwhelmed and compromised. According to Rawls, once ideal theory is completed, non-ideal theory can be set out by reference to the ideal. For instance, once we find ideal principles for citizens who can be productive members of society over a complete life, we will be better able to frame non-ideal principles for providing healthcare or some other programs to citizens with serious illness and disabilities.

the greatest happiness or good for all? Rawls answer comes in three parts. Firstly, the utilitarian principle, assumes and wrongly at that, that the happiness of two distinct persons can be meaningfully counted. Secondly, the principle, lumps together individuals who have separate goals to pursue, and treats a group of many as if it were a single person. Thirdly and more importantly, his two principles of justice promote a social scheme that is consistent with our 'considered moral judgments': they respect our equal moral status as rational autonomous persons by advancing the bases of self-respect and guaranteeing equal liberties and opportunities, as well as economic resources.

One fundamental issue that arises in connection with Rawls' project in moral and political theory concerns the issues of whether or not it violates the thesis of individualism. Many think it does. We encountered the thesis in chapter one: the view that we may neither collapse a person's conception of the good with those of others nor compel anyone to accept the principles of social relationships. The principles of justice, for Rawls, are indexed to the standpoint of the least advantaged chooser. Contouring the principles of justice to reflect the interests of the least advantaged chooser is to reflect the condition and interests of a particular group, in this case the worst-off group. Because the principles reflect the condition and interests of the least advantaged member of society, one can reasonably ask if in doing so we do not sacrifice the condition and interests of others. Put in a different way, do the principles—which are contoured by the perspective of a particular group in society—not promote and advance the condition and interests of

the least well-off group at the expense of the condition and interests of the better-off group?

2.2 Social Justice and the Need for it in Democratic Society

From the outset, Rawls' project in moral and political theory has been guided by the question, what is "the most appropriate moral [conception of justice] for a democratic society?"[35] To this end, the problem of the social contract, in his view, is fundamentally a question of social or distributive justice. Rawls formulates this question within the general rubric of the larger philosophical question: "What is the most just and feasible arrangement of basic social institutions that realizes the core democratic values of liberty and equality of all citizens?" Or put simply, "How do we create a society that satisfies the basic demand of social justice, i.e. that respects the equal moral status of persons?"

The answer to the fundamental question of social justice, Rawls believes, is found in the hypothetical social contract, which yields what he calls the abstract and general principles of justice in ToJ. Throughout the development of his moral and political ideas, Rawls' primary concern is to devise a form of association that embodies a just and equitable society. For Rawls, a society is just and equitable so long as it represents a fair system of social cooperation between individuals who are free and equal. He believes that JaF, with its rigor in advancing the principles of justice—or the fundamental moral principles—and mutual advantage, provides the theoretical underpinning for a just and equitable society.

[35] Ibid, p. xviii.

It is important to note that Rawls, unlike Gauthier as we shall see later, does not set out to provide a direct response to the general problem of rational compliance. However, his account has something to say to that problem. Rawls believes that a society is just and equitable if it represents a fair system of social cooperation between individuals. We expect that when people benefit from a scheme of cooperation they would agree to bear the costs necessary to maintain it. Only a fair system of social cooperation has the authority to command the voluntary consent of everyone. Because a fair system of social cooperation respects our equal moral status as rational autonomous beings, we could theoretically agree to it.

If the contract demonstrates for us how we come to choose the principles of justice, and if these principles match our considered moral judgments of what a just society ought to be, Rawls believes he would have effectively devised a theory that supplants the main alternative theories of justice, particularly utilitarianism. As a theory, utilitarianism, in all its various forms, balances benefits over losses and aggregates benefits for those affected by a situation. Consequently, it collapses together the separate lives, interests and conceptions of the good of individuals. This, according to Rawls is wrong, for in the pursuit of one's conception of the good or happiness, no one, and indeed no theory, should require one to suspend or sacrifice one's interest and lives for the sake of others.

Rawls' criticism of utilitarianism reiterates the importance of adhering to the thesis of individualism. Utilitarianism violates this thesis because it takes an individual's value to be a function of the value of that individual's contribution to

the common good. The value you have is not an inherent one. If this view is repugnant to us, the reason is perhaps that we share the view that we are valuable irrespective of the value we add to others' lives. But one wonders if Rawls' commitment to the thesis of individualism is not compromised by his approach that evaluates the principles of justice from the standpoint of the least advantaged member of society. And if it is, one wonders whether JaF is significantly different from utilitarianism. Isn't shaping the principles to reflect the interests of the worst-off group a different way of making one's value dependent upon the instrumental value of their contribution to others' good? And if so, how much different is this from utilitarianism that collapses together the separate lives and interests of individuals? I leave this worry for now.

Although the social contract is a device of representation, providing people reasons for acknowledging the impartial or neutral political perspective embodied in JaF, the fundamental moral principles, according to Rawls, are identified independently of the contract. This makes Rawls' approach essentially different from other accounts of moral contractarianism (such as Gauthier's) in which the principles of morality are the outcome of a bargain or the rational preferences of individuals in the original contract. In JaF, agreements do not create morality even though they may generate the principles of social relationships.

Justice, Rawls says, is the "first virtue of social institutions, as truth is of systems of thought."[36] That Rawls considers justice fundamental to society is evident, as we have seen, from the fact that he structures his moral and political project around the issues of what he calls the most appropriate moral conception of

[36] Ibid, p.3.

justice for a democratic society. Rawls' approach, which takes "to a higher order of abstraction the traditional theory of the social contract as presented by Locke, Rousseau, and Kant"[37] is very much Kantian. He provides the most famous restatement of a Kantian social contract theory in ToJ. In this book, Rawls, like Kant, presents the arguments for the social contract not as an explanation of the origin of political authority but as a way of describing and justifying a form of social and political association. Morally, Rawls agrees with Kant that the main aim of a moral philosophy is to set out to seek or discover fundamental moral principles which are to decide the basic structure of society. Rawls claims that the principles of justice are fair so long as the conditions under which they are discovered are fair. If the conditions under which they are discovered are basically fair, justice proceeds out of fairness.

Rawls' procedure in ToJ is to seek out first the fundamental moral principles that will govern social institutions, or what he calls the basic institutions of society, and then apply them to particular practices. To this extent, the principles of justice are not guidelines to regulate individual morality; rather they are schemes meant to regulate the basic institutions of society, which have the most extensive effects on the prospects of individuals. Rawls writes:

> Many different kinds of things are said to be just and unjust: not only laws, institutions, and social systems, but also particular actions of many kinds, including decisions, judgments, and imputations. We also call the attitudes and dispositions of persons, and persons themselves, just and unjust. Our topic, however, is that of social justice. For us the primary

[37] Ibid, p. xviii.

> subject of justice is the basic structure of society, or more exactly, the way in which the major social institutions distribute fundamental rights and duties and determine the division of advantages from social cooperation. By the major institutions I understand the political constitution and the principal economic and social arrangements.[38]

If the principles are fair and just, then particular practices or actions are fair and just to the degree they conform to just institutions. The substance of Rawls' claim is that a democratic society is just when its institutions are guided by fundamental moral principles that rational persons would agree to from a position of moral equality. And individuals and their actions are just insofar as both conform to the demands of just institutions. These principles, Rawls says, are deeply implicit in ordinary moral awareness and are evidenced by our most considered moral judgments. The bases for the principles of justice are provided by practical reason in addition to certain psychological tendencies of human nature and the capacities for sociability.

Rawls says his interest is in social or distributive justice, i.e. justice as it concerns the way in which the major social institutions distribute fundamental rights and duties and determine division of advantages from social cooperation. The principles of social justice, according to Rawls, are to apply to deep inequalities occasioned by the genetic history of people, and by being born into different positions. Generally, these inequalities are accentuated by institutions of society in the way they parcel out rewards and benefits in ways that often favor those who have better starting places.

---

[38] Ibid, p.6.

To understand Rawls' project in moral and political theory we need to approach it through the various simplifying assumptions that he makes. First, Rawls limits consideration of social justice to rational persons, that is, normal, fully cooperating members of society.[39] Second, he omits from his discussion issues about criminal justice, and justice of the laws of nations and of relations between states. And finally, he focuses on the main institutions of society, or what he calls the "basic structure" of society, such as the interconnected system of rules and practices that define the institutions of property and family, legal procedures and the system of trials, the political constitution, the laws and convention which regulate markets and economic production and exchanges.

According to Rawls, the question of justice arises because we are in what Hume calls the "circumstances of justice."[40] These are circumstances of moderate scarcity of essential resources in which cooperation is necessary for individuals if they are to individually meet their needs. Because individuals live in conditions of moderate scarcity, or in circumstances of justice, disputes arise over costs and benefits. The idea of justice is inapplicable in situations of scarcity or abundance. In scarcity, one's survival depends on depriving someone else the means of survival. In such a zero-sum situation, an individual can hardly be judged unjust just because she appropriates the only available means of survival. In condition of abundance, on the other hand, everyone has as much of what they desire, thus there could hardly arise any dispute that would raise issues of justice. If X has what Y wants, why should Y dispute with X about it if Y could get another just like it,

[39] Rawls excludes the mentally deficient, children from the contract situation.
[40] Ibid, see section 22 for a detailed discussion of the circumstances of justice.

without any difficulty? Rawls' decision to limit considerations to circumstances of justice makes his account, for the most part, realistic. For individuals situated in circumstances of justice are neither moral saints nor perfect altruists; neither are they natural sinners nor rational egoists. They are humans *qua* humans, given their nature, under normal conditions of social life.

Rawls thinks it is necessary to apply JaF to what humans are capable of[41] because a moral conception of justice need to be feasible and stable. In addition, the realization of justice as fairness ought to promote in people an unfaltering will to act justly or do justice and a disposition to uphold just institutions (as that conception defines them individually). Rawls thus shows his awareness of the pervasive problem that instability creates for social contract theory. Thus, to the extent that Rawls believes that the fundamental moral principles are to govern the basic structure of society and to the extent that the basic structure of society shapes the lives of citizens and the pursuit of their conception of the good he agrees with the general view that constraints are necessary foundation for a just and equitable society.

2.3 The Hypothetical Contract and *Justice as Fairness*

Rawls' entire project can be parsed out into four parts. The first is the conditions under which the hypothetical agreement or contract is to take place. Second, the reason why Rawls believes we should think of justice in terms of fairness, that is the reason why he thinks the principles that arise from the contract satisfy the demand of an equitable and just society. The third is the rationale for supposing

[41] This is a variation of Kant's "ought implies can" dictum.

that his principles of justice would be chosen under the conditions specified. Finally, the justification of the argument that this shows that his principles of justice are the correct principles of justice, at least for modern political and democratic regimes.

In choosing the principles of justice from a position of equality and partial ignorance, agents in JaF are aware what these principles are for: they are to govern the basic institutions of society. Given that they are guided by the capacity for a sense of justice, they know that it is appropriate for them to follow through on the principles that have been chosen. This capacity for a sense of justice is therefore fundamental, because it is an integral constituent of what is it that motivates them to comply with whatever principles that they happen to choose in the original position.

2.3.1 Conditions under which Hypothetical Agreement takes Place

The first part of Rawls' project describes the general condition of contractors and the condition under which agreement is to take place. Rawls identifies the general condition with the "original position": a hypothetical situation developed by Rawls to replace the state of nature of classical accounts of social contract theory. The original position is designed to be a fair and impartial standpoint that is to be adopted in our reasoning about fundamental principles of distributive justice. The original position idealizes agents by situating them in certain special conditions; conditions that Rawls says "are the ones that we do in fact accept. Or if we do not,

then perhaps we can be persuaded to do so by philosophical reflection."[42] There are two parts to the conditions in the original position. These are conditions of knowledge and conditions of ignorance.

Concerning conditions of knowledge, in the original position, agents know they possess general, uncontroversial knowledge. They are situated equally, i.e. they posses the same rights and moral powers, and they understand that their society is situated in the circumstances of justice. They know that the principles of justice they select are to be public and are to effectively regulate the basic structure of society. They also know that they are capable of acting on a conception of justice, specifically, they are motivated by the capacity for a sense of justice. Closely related to the capacity for a sense of justice is what Rawls calls the 'constraints of finality,' by which he means that agents are willing to bear the 'strains of commitment' once the contract is concluded. Both are crucial to Rawls' argument that contractors would voluntarily comply with the principles that emerge from the contract. Since they know that the contract is made in good faith, they do not seek to annul it or have it revoked just because they got the wrong end of the stick or things turn out badly for them.

The capacity for a sense of justice is one of the two moral powers that Rawls says define the conception of moral persons. The other one is the capacity for a conception of the good, that is, to form, revise, and rationally pursue a

[42] Rawls, ToJ, p.19.

rational plan of life. Both moral powers, according to Rawls, are the grounds for full autonomy or rational and moral agency.[43]

Outside the conditions of knowledge, there are conditions of ignorance. The agents are ignorant about certain things, or as Rawls puts it they are placed behind a "veil of ignorance,"[44] which makes them unaware of their personal and particular circumstances. They do not know their place in society or their class position. They are ignorant of their social status, their gender, their conception of the good, their religion, their race, and more importantly, they are ignorant of their special psychological propensities and their possession of natural assets—their abilities, talents, strengths and level of intelligence. They do not know the particulars of their society, its economic and political situation, its level of civilization or culture, or the generation to which they belong. The advantage of this, according to Rawls, is that it prevents the agents from contouring principles of justice to the benefits of a particular social class or circumstances.

There is something troubling about Rawls' veil of ignorance. It would seem that decision- makers deprived of knowledge about themselves, their interests and

[43]Rawls revises this view (Kantian constructivism) in *Political Liberalism* (New York, Columbia University Press, 1993) following his recognition of the tension that is implicit in his stability and congruence arguments (as contained in the ToJ), and the criticism that the lack of moral neutrality of his moral conception of the person undermines the feasibility of justice, and hence creates the problem of stability. Rawls' moral conception of the person construes these moral powers as constituting our nature as moral beings, or necessary conditions for moral agency, such that realizing them enables us to realize the supreme human good of autonomy or the good of free and equal rational beings. In *Political Liberalism*, he then presents the moral powers not as a moral conception of the person but as a political or freestanding conception of the person, as empirical conditions for achieving the advantages and benefits of social cooperation, without which individuals could not comply with the duties, or take advantage of the rights, of democratic citizens, whatever their conceptions of the good is. Rawls states the general question addressed in *Political Liberalism* in this way: "How is it possible for there to exist over time a just and stable society of free and equal citizens, who remain profoundly divided by reasonable religious, philosophical, and moral doctrines?" p.4.

[44] See Rawls, ToJ, section 24.

abilities, their circumstances and their conception of the good would be unable to make any meaningful choices or decisions regarding whether they value liberty and how society ought to be. Furthermore, libertarians and some contractarians have argued that screening information from decision-makers ignores the history of legitimate acquisitions. Information about the starting or bargaining position of decision-makers, especially property rights, they argue, is relevant to any theory of justice or social contract insofar as the acquisitions are *legitimately* acquired.

Rawls is not unaware of this and in response he makes some additional assumption about agents and what information that the veil of ignorance filters out: the assumption that decision-makers have certain motivations and that possess some form of, or a thin conception of the good. This additional assumption enables Rawls to say that although decision-makers do not know their conception of the good, and their special psychological propensities, they however, do know that they have certain motivations that gravitate towards some conception of the good (their *summun bonum*). By this Rawls means that they do possess a 'thin, formal conception of the good' that they know would be consistent with any of the richer, special conceptions of the good that might be their actual one, once the veil of ignorance is lifted.

What does Rawls mean by the statement that 'agents possess a thin, formal conception of the good?' What he means is that they know that they want 'social primary goods,' i.e. what all "persons need in their status as free and equal citizens, and as normal and fully cooperating members of society over a complete life."[45] The social primary goods are what people rationally want, whatever their special

[45] Ibid, p. xiii.

conception of the good, or whatever else they want. These are liberties, opportunities, income, wealth, and the social bases of self-respect. Because agents in the original position are rational, they prefer more of the social primary goods to fewer, and they take the most efficient means to achieve whatever ends they happen to have. And since they are 'mutually disinterested,' that is, they are concern only about their own interests, and not envious, they care less about the social and economic status of others.

Regardless of the additional assumption that Rawls makes, his agents are still far from determinate. His social contract account seems to be so abstract that it can only yield abstract principles of justice. The problem with the abstract principles of justice in JaF is that they cannot effectively guide the choices of goods, values and actions. Communitarians argue, for example, that the standards of justice must be found in forms of life and traditions of particular societies. And because the standards of justice must be grounded on the habits, practices, institutions, beliefs and traditions of actual people living in specific times and places, any theory of justice that abstracts away from these forms of life and traditions of particular and determinate individuals and societies, as Rawls does in JaF, cannot effectively govern social practices and political society.[46]

For different reasons, Gauthier also criticizes Rawls for leaving out or abstracting away from the preferences of determinate individuals. In choosing the

[46] See for example, Alasdair MacIntyre, *Against the Self-Images of the Age*, Notre Dame, University of Notre Dame Press, 1978, chs. 18-22, *Whose Justice? Whose Rationality?*, Notre Dame, University of Notre Dame Press, 1988, ch. 1; Charles Taylor, *Philosophy and the Human Sciences: Philosophical Papers 2*, Cambridge, Cambridge University Press, 1985, ch. 1; Michael Walzer, *Spheres of Justice*, Oxford, Blackwell, 1983, p.8.

principles of justice, agents in JaF are prevented from knowing what their preferences are in order to prevent them from contouring the principles of justice to their particular circumstances. Gauthier objects that agents deprived of their preferences cannot meaningful choose any principle of justice that is to regulate individual behavior. Gauthier's moral contractarianism which, as we shall see in later chapters, grounds bargaining on determinate individuals and their preferences.

Rawls is right to claim that an adequate social contract theory is one that appropriately defines contractors in the original position. However, if the principles of justice that regulate social and political institutions, markets, economic production and exchanges are to be motivationally efficacious or command the willing compliance of decision-makers, then they would have to speak to their considered preferences. Given that such preferences are crucial to securing the commitment of contractors, any contract theory that omits them sets itself up to be attacked by Hobbes' Foole. This takes us to the second part of Rawls' project: why we ought to think of justice in terms of fairness.

2.3.2 Why we should think of Justice in terms of Fairness

Rawls believes that the principles of justice that are chosen behind the veil of ignorance embody a commitment to liberal neutrality and a commitment to justice construed as fairness to all. The reason why we should think of justice in terms of fairness, according to Rawls, is that only such a conception of justice meets our considered moral judgment of fairness, i.e. represents the core democratic values of liberty and equality of all citizens. A principle of justice is fair to all when that

principle takes an unbiased and neutral standpoint regarding the conceptions of the good of all. When the principle of justice is not skewed by individual characteristics and circumstances, it treats all citizens as equals. When it recognizes the equal moral status of all, it takes an appropriate stance towards their interests.

To illustrate, consider a society that satisfies the condition of the circumstances of justice, say one that is equally divided between the rich and the poor. How should we proceed in order to resolve issues of justice?[47] One way is to try and get all to agree on terms to regulate the economic situation of the society. But this would be impossible if all were to decide with full knowledge of where they belong in the social and economic spectrum. The result would be that there are no terms to which everyone literally would agree that would satisfy a complete conception of justice.

We can reasonably expect that the rich would be opposed to taxation, preferring a laissez-faire system that would enable them to keep most, if not all, of their income, while the poor would approve of taxation of the rich, in order to provide for welfare benefits like healthcare, education and other social programs. Part of the aim of a theory of justice is to resolve such disputes. It cannot be achieved by extracting principles of justice from a poll of the rich and poor who are biased by their social and economic status and preferences.

The best way of resolving the dispute, according to Rawls, would be to put the rich and poor people behind a veil of ignorance that blinds them to their economic status and preferences. Rawls supposes that if people do not know what

[47] The substance of this example is drawn from Jonathan Wolf's *An Introduction to Political Philosophy*, revised edition, Oxford, Oxford University Press, 2006, p.154.

their status is in terms of talents, abilities, intelligence, conception of the good, etc, or where they belong in the spectrum of the natural lottery, they cannot be biased in their choice of principles of justice. Rawls believes that justice requires impartiality and that impartiality can best be modeled both by construing agents in the contract situation as free and equal and by assuming ignorance, hence his claim that the condition of ignorance in the original position leads to the correct principles of justice.

Rawls here adopts the Kantian perspective—captured later by Thomas Nagel's "The view from nowhere"[48]—of not skewing the moral law to one's circumstance, namely, of not making exceptions to one when willing a particular maxim into a universal law. Behind the veil of ignorance, every person is presumed to be equally rational. Since all contractors adopt the same method for choosing the principles of justice and since they are equally rational, they will occupy the same standpoint: that of the disembodied, rational, universal human, which is the same with the point of view of the least advantaged member of society. Therefore, every person who considers justice from the standpoint of the original position, properly situated behind the veil of ignorance, would reach the same result, namely: agree upon the same principles of justice.

[48] In *The View From Nowhere*, Nagel makes a distinction between the two points of view from which we see the world. The first point of view is the 'subjective point of view,' namely, the view of our conscious selves. We think of the world, in terms of our experience. The second point of view is the 'objective point of view' or 'the view from nowhere.' We think of the world in terms that transcend our experience, namely, independently of both our viewpoint or any other particular viewpoint. It is the second point of view that Rawls adopts in JaF. See Nagel, *The View From Nowhere*, New York, Oxford University Press, 1989, chapter 5 (knowledge).

2.3.3 Reason why the Principles would be chosen under the Specified Conditions

I now come to the third part of Rawls' project: his claim that specifying the original position the way he did leads to his two principles of justice. Rawls believes that he has set up not just an abstract procedure in JaF, but also an inherently fair one. Because of the fairness of the procedure, he says, the principles of justice that would be chosen by means of this procedure would be essentially fair principles. The principles that Rawls says would be chosen in the original position are:

1. Each person is to have an equal right to the most extensive total system of equal basic liberties compatible with a similar system of liberty for all.
2. Social and economic inequalities are to be arranged so that they are both
    a. to the greatest benefit of the least advantaged…and
    b. attached to offices and positions open to all under conditions of fair equality of opportunity.[49]

The two principles of justice—the liberty principle (or the equal rights principle) and the social and economic principle (or the principle of wealth and opportunities or the principle of permissible social and economic inequality)—Rawls says are specific in their content. He describes them as special formulation of a more general conception of justice. The general conception of justice is stated as:

[49] Rawls, ToJ, p.266.

> All social values [social primary goods]—liberty and opportunity, income and wealth and the social bases of self-respect—are to be distributed equally, unless as unequal distribution of any, or all, of these values [goods] is to everyone's advantage [or the advantage of the least favored].[50]

In the more general conception of justice, individual liberties and economic benefits are thought of as equally important. Liberty may be sacrificed for greater economic benefits. Notice that this view changes in a very important way in the special conception of justice that incorporates the two principles of justice.

The significance of Rawls' general conception of justice is that justice is only completely operational under sufficiently advanced social conditions. Rawls does not define what the precise conditions are for the two principles of justice. However, the idea that principles of justice can only fully operate in the presence of sufficiently advanced social conditions is compelling. The insightful idea is that before reaching a certain stage of development (say, industrialization), a society can reasonably adopt a quasi-utilitarian standpoint of sacrificing individual liberties for greater economic benefits. In other words, before attaining a certain stage of social and economic development it is morally permissible for a society to violate individual liberties in order to bring about that stage of development that will enable the principles of justice to be fully operational.

Rawls' two principles of justice imply that no matter what one's position is in society after the veil of ignorance is removed one will have the necessary liberties and resources that would enable one exercise one's rational capacities and

[50]Ibid, p.54.

to pursue whatever conception of the good one happens to have. The principles guarantee that people, no matter where they are in the spectrum of the natural lottery, are treated as free and equal citizens. They also enable individuals to maintain their self-respect, which Rawls says is "perhaps the most important primary good."[51] Self-respect ensures that people have confidence in their own abilities and sense of their own value, for when people,

> feel that [their] plans are of little value, [they] cannot pursue them with pleasure or take delight in their execution…. Therefore the parties in the original position would wish to avoid at almost any cost the social conditions that undermine self-respect.[52]

The principles of justice that emerge from the contract situation, Rawls says, do not result in a zero-sum game but a win-win situation for everyone. Justice as fairness satisfies the condition of mutual advantage because it ensures that no one depends on others for the protection of their interests and that all are socially and economically self-sufficient. Because citizens are not subservient to the will of another, they stand to one another as equals and not as superiors and subordinates. And because the principles ensure the most important social primary good, it would command the assent of everyone and guarantee greater overall stability. Achieving and maintaining the bases of self-respect is therefore crucial to upholding the standard of equality and fairness. Whenever the basic structure of society deviates from the ideal of equality, we have socio-economically imposed unfairness. A society that fails to treat some of its members as equals is one that

[51] Ibid, p.386.
[52] Ibid, p.386.

either restricts their liberties (e.g., their right to life, or freedom of conscience, or expression) or permits them to grow up in destitution.

Principle 1, the liberty principle, is a principle of strict equality and applies generally to the constitutional structures and guarantees of the political and legal systems. Principle 2 is the principle of social and economic wealth. Principle 2(a) is the difference principle, and 2(b) is the fair opportunity principle, and they both apply to the operation of the social and economic systems. These principles, according to Rawls, are lexically ordered such that the liberty principle takes precedence over the fair opportunity principle, which takes precedence over the difference principle. The lexical priority of these principles means that liberties cannot be sacrificed for equal opportunity and equal opportunity cannot be sacrificed for economic welfare or some other values that we may hold. Liberties can only be limited for the sake of maintaining other liberties, and fair opportunities can only be limited by liberties and fair opportunities, not by economic welfare.

The liberty principle, which guarantees 'basic' rights and liberties, is straightforward and familiar from much liberal thought. The basic liberties guaranteed by the principle are:

> Political liberty (the right to vote and to be eligible for public office) together with freedom of speech and assembly; liberty of conscience and freedom of thought; freedom of the person along with the right to hold (personal) property; and freedom from arbitrary arrest and seizure as defined by the concept of the rule of the rule.[53]

[53] Ibid, p.53.

The basic liberties are *inalienable*, Rawls says, because they are essential to a person's sense of self-respect and to the full development of the capacity for a sense of justice. These rights cannot be traded away because they are what are needed if people are to exercise their moral powers or to be defined as free and equal persons.

The first principle guarantees negative liberty of the citizens. They are to be free from unwarranted state interference with their lives. Rawls draws on the traditions of classical liberalism that in contemporary political discussions are argued for most forcefully by libertarians, such as Nozick in *Anarchy, State and Utopia*.[54] The liberty principle is fairly uncontroversial, for the idea of individual liberty has become so entrenched in our cultures that practically everyone accepts it in some form.

Rawls thinks that because of the status that he accords liberties, this makes their protection better under JaF than, for example, under libertarian schemes that champion unlimited liberties of people under unrestricted regimes of contract and transfer but leave them unsecure and unprotected. Another significant area where JaF differs from libertarian views is in the area of natural or presocial property rights as espoused by Nozick and Gauthier. For Nozick, property rights are natural and absolute. They determine the legitimacy or illegitimacy of what anyone, including the State, can do.[55] For Gauthier, property rights are natural in the sense that they are necessary for social interactions and the emergence of the market.

---

[54] Nozick, *Anarchy, State and Utopia,* New York, Basic Books Inc., 1974. See especially chs. 3 and 7.

[55] Nozick views individual rights as side constraints. He boldly announces the constraining nature of rights in the opening pages of his influential work, *Anarchy, State, and Utopia* this way:

By contrasts, property rights, for Rawls, do not enter into the protection of the liberty principle as do the other basic rights. Even though property rights are institutional, JaF leaves them to be regulated by the difference principle on the basis of the ideal of 'reciprocity' which requires that the position of everyone be advanced in society in such a way that those better off do not achieve their gains at the expense of the less advantaged and fortunate. By the ideal of reciprocity, Rawls means that

> the institution of property is justly ordered when it is part of a social and economic system that specifies property relations so as to make the worst off class better off than they could be under the institutions of any feasible alternative economic system (subject to the conditions that equal basic liberties and fair opportunities are always maintained).[56]

The second principle, the principle of social and economic benefits, is a principle of positive liberties. This principle, according to Rawls, "applies in the first approximation, to the distribution of income and wealth and to the design of organizations that make use of differences in authority and responsibility, or chains of command."[57] The fair opportunity principle is a principle against formal discrimination. It is also a principle that so to speak, requires leveling the playing field for all. The principle requires that whatever individuals' status or talents may be, they should be provided the same opportunity to try to develop their natural talents and abilities so as to compete for jobs and positions without handicaps and

---

"Individuals have rights, and here are things no person or group may do to them (without violating their rights). So strong and far-reaching are these rights that they raise the question of what, if anything, the state and its officials may do," p. ix.

[56] Samuel Freeman, "Introduction: John Rawls – An Overview," in *The Cambridge Companion to Rawls*, Samuel Freedman, (ed.), Cambridge, Cambridge University Press, 2003, p.7.

[57] ToJ, p. 53.

obstacles arising from a deprived background and natural history. This is a positive interpretation of equality of opportunity, which is different from the negative interpretation that requires only the absence of barriers to places or competition for positions in the social and economic ladder.

According to negative understanding of equal opportunity, it is good to proscribe intentional discrimination, for no one who is qualified ought to be excluded from profession and positions because of their sex, religious convictions, and race. To do so is to unjustly penalize an individual on grounds that are morally irrelevant and arbitrary in the worst sense. Any society that allows such a thing is not only immoral but also unjust. Rawls rejects negative equality of opportunity, which he considers unstable, because it does not provide for complete equality of opportunity. Since people are neither responsible for their natural history nor for the socioeconomic status of the family into which they are born, we need a system that ensures that individuals can acquire the necessary training and background to develop their natural talents and abilities so as to compete for formally available opportunities and open positions. The principle of fair opportunity thus requires not only that no one be discriminated against on the basis of their sex, race and religious convictions but also that measures are taken to ensure that those whose starting place in society is less favorable have an equal chance to achieve an important position in society.

The difference principle is strongly egalitarian and is meant to deal with structural inequalities that affect statistically great numbers of people in the different socioeconomic strata. It is not meant to deal with inequalities that arise

owing to the decisions people make, that is "inequalities among individuals that will inevitably arise as people make choices and interact, and succeed or fail in their efforts, in the context of any socioeconomic structure, however just."[58] In other words, the difference principle supplements the fair opportunity principle just as the negative equality of opportunity is complemented by the positive provisions of the fair equality of opportunity principle.

Rawls' argument is that undeserved inequalities would continue to arise even under a regime of fair equality of opportunity. To the extent that it can be realized, fair equality of opportunity guarantees only that people of equal natural ability will have roughly equal chances to flourish. But this is a bit misleading because people are not equal in natural ability. People are naturally or genetically different and these differences will continue to affect the advantages and benefits they gain from interaction with the socioeconomic arrangement. But this too is morally arbitrary, for people are no more the architect of or responsible for their genetic endowments than for their race, sex or the economic status of their parents.

Accordingly, a just society will blunt these undeserved differences in genetic endowments and benefits to the extent that it can do so without harming those whose arbitrary penalization it is most concerned to rectify, namely, the worst-off group. And that is exactly what the difference principle does in deeming structural inequalities to be unfair and intolerable for any decent society with the resources to prevent them. A society is just to the extent it pushes toward the elimination of inequalities that are not necessary for the provision of maximum

[58] Thomas Nagel, "Rawls and Liberalism," in *The Cambridge Companion to Rawls*, Samuel Freedman, (ed.), p. 71.

benefit to the worst-off group. Inequalities can only be justified if the institutions that make up the basic structure of society are the most effective and available in achieving an egalitarian end, namely that of making the worst-off group in the society as well off as possible. The difference principle and the fair opportunity thus prove to be closely related to the liberty principle—all aim towards provision of a 'social minimum'; a basic social entitlement to enabling resources required in the exercise of basic liberties. For without the social minimum that is provided for by the social and economic principle, the basic liberties are valueless and empty; mere formal protections worth little to people who are impoverished and without the means to take advantage of their liberties.

The social and economic principle is based on the idea that negative liberty by itself is inadequate. Given the existence of natural inequalities, a great number of individuals will be unable to exercise their negative liberty in any meaningful sense unless the contingencies of nature are taken into account. The negative liberty of the first principle would seem of little potential value for someone who cannot get any meaningful job because they lack the resources to develop their talents. The fair opportunity principle ensures that no person is in that position. And the negative liberty of the first principle does not empower someone who lives on the street and has to beg to survive. The difference principle ensures that no one is in that situation.

The liberty principle, Rawls says, would be chosen in the original position because contactors would prefer a society were every one has an equal right to extensive liberties to a society with unequal or limited liberties, or even slavery,

notwithstanding whatever economic benefits these societies might produce. To refuse the extensive liberties that the liberty principle makes possible is to endorse in some form a social scheme that encourages slavery. Since the veil of ignorance prevents the agents from knowing what their status is in society, the rational and automatic choice is for them to choose in a way that they would not generally be disadvantaged. Hence, the social and economic principle would be rationally chosen. To reject the provisions that the principle makes available for the maintenance of the bases of self-respect and pursuit of one's conception of the good is to endorse a social scheme of grave inequalities, whereby some people may have to beg on the street to survive.

From the foregoing, it is indeed clear that Rawls has constructed what perhaps can be considered the most elegant and abstract version of contractualism. JaF is extremely ideal for a number of reasons. There is first the point about indeterminate agents, who as we have seen, lack personal knowledge about themselves in adopting the principles of justice in the original position. Secondly, there is the issue about the hypothetical agreement. Rather than demonstrating that rational agents *would* agree to a contract to establish society, Rawls' social contract theory demonstrates instead what they *must* be willing to accept as rational persons in order to be constrained by the principles of justice if they are to live in a well-ordered society. The abstractness of JaF is most evident in the principle of choice that Rawls says is employed by agents in choosing the principles of justice in the original position. I now turn my attention to this principle of choice.

2.3.3.1 The Maximin Principle, Rational Choice, and the Principles of Justice

There are different strategies of choice that agents can adopt in managing risks in situations of uncertainty. Given the degree of uncertainty that Rawls' contractors face, he argues that the rational strategy open to them is the "maximin principle of rational choice." The maximin principle instructs that the minimum should be maximized, and it stipulates that agents, in the original position, ought to choose in such a way that the worst possible outcome is as good as possible, that is to say, the worst possible outcome of the social scheme they choose is better than the worst possible outcome of *any other* social scheme.

Rawls' employment of the maximin rule, as we shall later see, contrasts with a moral contractarian approach of minimax relative concession that Gauthier defends. Rawls imposes a strong egalitarian constraint on the pursuit of individual interests in society by arranging social and economic inequality from the standpoint of the least advantaged member of society such that an individual may benefit just in case it maximizes the minimum benefit. By contrast, a moral contractarian approach considers social and economic inequalities from the standpoint of every decision-maker, that is, the Archimedean chooser. In choosing principles, they choose as if they were each bargainer.

Proponents of strict Bayesian decision theory have criticized Rawls' maximin rule, labeling it an irrational rule of choice. According to John Harsanyi, a leading Bayesian decision theorist and moral contractarian, Rawls "makes the technical mistake of basing his analysis on a highly irrational decision rule, the

maximin principle, which [has] absurd practical implications."[59] Harsanyi claims that the only rational principle of choice that agents can adopt in the conditions that Rawls describes is one that satisfies 'the principle of insufficient reason' according to which one picks the choice that has the greatest benefits, i.e. maximizes EU no matter what is at stake. We can formulate Harsanyi's criticism of Rawls' maximin principle in the form of a weak dilemma:

> **P1:** Rawls' agents are either irrational or rational.
> **P2:** Agents cannot maximin under conditions of uncertainty, and to suppose they do is to assume that they are irrational.
> **P3:** But Rawls assumes that his agents are rational.
> **Conclusion:** Therefore, they cannot maximin.

Since Rawls assumes that his agents are rational, and since he thinks they maximin, how does he respond to the dilemma posed by Harsanyi? How does he show that the maximin rule is not irrational? If, we suppose that the maximin rule is irrational, to what extent can it be said to be? Stated differently, are there circumstances where the maximin rule can be said to be a rational strategy of choice? And does the fact that there exist situations where the maximin rule is irrational lead to the conclusion that it is always irrational?

The maximin principle, according to Rawls, is (1) "in general a suitable guide for choice under uncertainty" and (2) "holds only in situations marked by certain special features."[60] Given Rawls' view that the maximin rule is applicable in situations that satisfy specific characteristics and conditions, he seems to accept

[59] Harsanyi, "Morality and the Theory of Rational Behavior," in *Utilitarianism and Beyond*, Amartya Sen and Bernard Williams (eds.), Cambridge, Cambridge University Press, 1982, p.47.
[60] ToJ, p.133.

the point that although it is not always rational to apply the maximin as a rule of choice, there are circumstances where we can and ought to apply it. This seems intuitively evident. We can say, for example, that, in general, one ought not go over the speed limit, but still acknowledge that there are circumstances that one should, say, in an emergency.

This sentiment seems to be echoed by Samuel Freeman. According to Freeman, to be able to determine the rationality of a decision rule, it is necessary to consider what is at stake, what the content of the decision is. If we assume that the content of the decision is a

> choice between bets on horses or voting on some minor legislation insufficient reason might be the rational strategy if you are ignorant of relevant information, for you will have the opportunity to play again and recoup your losses. If *however, a person's life is at stake, or all future prospects, then it seems a different matter entirely*.[61]

If what is at stake is something substantial such as one's future prospects or life, and one's options are unsupported by evidence, the rational strategy to adopt would seem to be the maximin and not one that maximizes EU. The principle of insufficient reason says that if we are confronted by options to which we cannot assign any quantum of probabilities, then the rational thing to do is to assign equal probability. But if one has no grounds for the assigning of probability to one's options, then it would also follow that one has no reason to assign equal

[61] Freedman, p.17, emphases are mine.

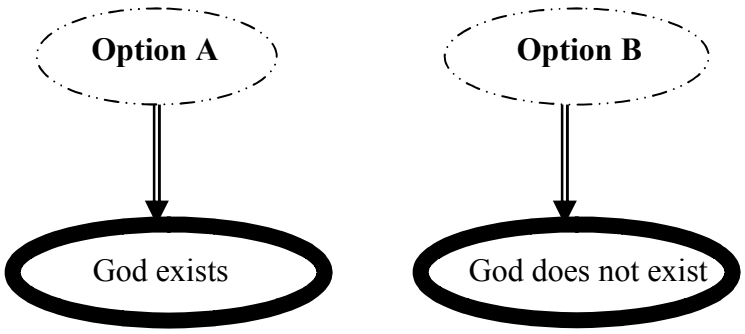
probabilities, especially when what is at stake is substantial or when there exists what Freedman calls "an acceptable alternative."[62]

This line of reasoning has been well articulated by Harold Jeffrey, whom Freedman quotes:

> If there is no reason to believe one hypothesis rather than another, the probabilities are equal…*to say that the probabilities are equal is a precise way of saying that we have no good grounds for choosing between the alternatives*.... The rule that we should take them equal is not a statement of any belief about the actual composition of the world, nor is it an inference from previous experiences, it is merely the formal way of expressing ignorance.[63]

To illustrate, consider the case of an agnostic philosopher who lives in Adejeland and is confronted by the following situation:

Figure 2.2.4a: **Choice and Sufficient Reason**



Let us call our agnostic philosopher Adejian. On the strength of the evidence before her, Adejian has no reason to believe either A or B; hence she is agnostic

[62] Ibid, p.17.
[63] Harold Jeffrey quoted in Freedman's "Introduction: John Rawls – An Overview", fn. 31, emphases original.

about the existence of God. Let us suppose that what we have before us is what William James calls a *forced option*,[64] or we can say that beginning from tomorrow sitting on the fence is no longer an option available to anyone in Adejeland. Since this is now the case, our agnostic philosopher would have to decide between A and B, and given that there is no new evidence for her to enable her chose one way or the other, whatever choice she makes (A or B) cannot be justified in reference to evidence or sufficient reason.

Adejian has more or less been thrown into a 'Buridan ass archetype like' situation in that like an ass placed between equidistant stacks of hays of equal size and quality she cannot make a rational decision to pick one option rather than the other. Unlike the Buridan's ass, she is not faced with starvation. However, their situations are similar in one relevant aspect—they both are forced to make a decision. And in making a decision, Adejian has no reason to assign probabilities to A or B. But since she is put in a position where she has to choose, it would be mistaken to say her choice is irrational, whatever her choice is.

For some the choice between A and B may seem 'trivial', but let us say that for Adejian, it is 'momentous,'[65] one that have enormous consequences to, say, one's life. What would be the rational choice for our agnostic philosopher? Imagine that the following are the possible distributive schemes facing her :[66]

---

[64]William James makes a distinction between an *avoidable option* and a *forced option*, in "The Will to Believe," which was first published in the *New World*, June 1896. An *avoidable option* for James is one that does not require a decision; in a *forced option*, there is no standpoint outside the decision.
[65] See James' "Will to Believe" for a distinction between *trivial* and *momentous options*. Whereas in a *trivial option* the opportunity is not unique, the possible consequences are insignificant, or the decision is reversible; but in a *momentous option*, the decision is about a unique opportunity, the possible consequences are significant, and the decision is irreversible.
[66] This scheme of distribution is a variant of the one Rawls consider in ToJ, p.133.

Figure 2.2.4b: **Employing Maximin in Situations of Uncertainty**

	S^1	S^2	S^3	*Sum EU (assuming equal probability)*
Option A^1:	10:	7:	-2	**7.5** (*10* x .5 + *7* x .5 + *-2* x .5)
Option B^1:	6:	5:	2	**6.5** (*6* x .5 + *5* x .5 + *2* x .5)

Let the numbers represent the quality of life in three possible states or circumstances ($S^{1,}$ S^2 and S^3). Let us suppose that the higher the number the higher the quality of one's life. If we assign equal probability to options $\mathbf{A^1}$ and $\mathbf{B^1}$, option $\mathbf{A^1}$ would be chosen by Adejian if she employs the Bayesian principle of sufficient reason since the EU (**7.5**) of states 1 to 3 ($S^{1,}$ $S^{2,}$ S^3) of option $\mathbf{A^1}$ is higher or greater than the EU (**6.5**) of states 1 to 3 ($S^{1,}$ $S^{2,}$ S^3) of option $\mathbf{B^1}$. But if we assume that the circumstances of the choice that confronts her make it a momentous one, one that impacts significantly on her life, then it would appear rational for her to choose option $\mathbf{B^1}$ since no matter where she ends up ($S^{1,}$ $S^{2,}$ S^3) she would have at least *2* (rather than *-2* if she were to choose option $\mathbf{A^1}$). Moreover, since the circumstances of choice that confront Adejian make it not just a situation of uncertainty, but one where knowledge of probabilities is impossible, or at best extremely insecure, the maximin rule, according to Rawls, turns out to be a rational strategy in this case. In employing the maximin rule, agents secure greater benefits from society than they would under any other decision rule.

Unlike the maximin principle, other principles of rational choice like the maximax, or utilitarian principle that maximizes expected benefits, Rawls claims, involve taking chances that are too risky to be rational in these circumstances of

choice. If one decides to gamble and take some serious risk, given that there is no second chance, one is stuck and doomed if one loses or ends up in the lowest rung of the ladder. The circumstances of choice are radically uncertain and—given the constraints of finality—the principles chosen are final and irreversible. In view of the fact that choice in the original position is a once-only choice with no going back, i.e. it would not be replayed, Rawls argues that the rational principle to employ is the maximin. Stated simply, given that there is always the possibility that one will have the misfortune to end up very badly or worst-off, agents would choose to maximin and choose the social and economic principle.

Beyond this, the maximin rule, Rawls claims, shows what is wrong with aggregative theories like utilitarianism that lumps benefits. Aggregating benefits is to ignore the fact that people have separate lives and different conceptions of the good. By insisting against utilitarianism on the separateness of persons, Rawls carries forward Kant's theme of respect for persons and points to the importance of respecting the thesis of individualism. The maximin rule, according to Rawls, meets the demand of rationality, which requires that we respect persons as ends in themselves, which in turn requires that the fundamental moral principles be such that they can be justified to each person. However, even if we suppose that the maximin rule preserves the idea of the separateness of persons, it is far from clear that the difference principle emerging from the contract respects persons as ends in themselves.

If the standpoint from which justice or mutual advantage is evaluated is that of the least advantaged person, then the principle imposes a strong egalitarian

constraint on the pursuit of individual interests in society, by requiring that each person benefits if and only if the minimum benefit is maximized. Thus, Rawls' approach is not after all so fundamentally different from utilitarianism. Whereas JaF, in virtue of the difference principle, circumscribes justice by the maximization of the minimum benefit, utilitarianism circumscribes justice by the maximization of aggregate or total benefits.

In any case, the argument can be made that since the position of the least advantaged chooser is indeterminate, Rawls' two principles of justice cannot be rationally chosen. For rational agents to choose principles from the standpoint of the least advantaged person, they would have to choose as abstract agents. And because the agents are not determinate, the principles of justice that arise from the original position can hardly regulate society comprising of real and determinate persons. On the contrary, if we suppose that the principles of justice are selected from the standpoint of every determinate person, then the maximin rule is not a rational principle to adopt in choice situations, and the social and economic principle, it would seem, would not be rationally chosen.

2.3.4 The Justification of the Principles of Justice

This takes us to the final part of Rawls' project: that of justifying the principles as the correct principles of justice. For Rawls, the principles are justified if it can be shown that the steps leading to them and the hypothetical contract, i.e. every element of the hypothetical contract including the principles, are fair and just. And

if we can justify the principles we would have provided a reason why we ought to comply with them in situations where we are encouraged and tempted not to.

Every element of the contract, according to Rawls, is fair and justified just in case two conditions are satisfied. The first condition is that agreement in the original position is feasible. The second is that every element of the contract reflects relatively uncontroversial moral beliefs that we do hold or could be brought by philosophical reflection to hold. The first is fulfilled once agents in the original position are characterized in such a way that they can come to some agreement or other. The second is satisfied by means of the method of reflective equilibrium. Rawls believes his method fulfils both conditions.

The first condition, he says, is satisfied through the way he described the original position, understood in the light of the ideas of a political conception of justice and of an overlapping consensus, which jointly leads to agreement among the contracting parties.[67] A political conception of justice is 'freestanding' or independent insofar as it is not grounded on premises peculiar to metaphysical, epistemological, and general moral conceptions. And there is overlapping consensus when a freestanding political conception of justice is supported by disparate conceptions of the good such that those who affirm these conceptions can be rationally motivated to do what the political conception requires.

The second condition is fulfilled when the elements of the contract and principles match our considered moral judgments about justice in reflective equilibrium. If they do not, then we are to revise the constraints on choice in the

[67] See *Political Liberalism,* pp.29-35 on Rawls' discussion of how a political conception of the person affects the understanding of a political conception of justice and an overlapping consensus.

hypothetical contract until we arrive at a contract situation that yields principles that are in reflective equilibrium with our considered moral judgments about justice. These judgments, which may be judgments about general or specific moral principles or cases, may conflict in some way. When they do, we proceed by adjusting our various beliefs and ideas until they are stable or in equilibrium. Moral beliefs that are stable or in ideal reflective equilibrium provide consistent practical guidance, and they, according to Rawls, describe the fundamental principles of our sense of justice.

Furthermore, the device of the contract and the principles are in reflective equilibrium when they match with certain background theories that themselves contain moral beliefs. These are theories or fundamental beliefs about the nature of persons, and the role of morality or justice in society. Recall that Rawls holds that putting agents behind a veil of ignorance ensures impartiality, and this, he claims, reflects our beliefs that systemic forms of discrimination are wrong. He argues that the principles of justice conform to our knowledge of human nature, especially the human desire for self-respect and the natural moral capacities to reciprocity, that allow the contractors to recognize and respect the legitimate interests of others while at the same time freely promoting their own good.

But doesn't the use of the method of reflective equilibrium to circumscribe acceptable moral beliefs and the device of the contract situation undermine the justification of Rawls' approach to justice? Grounding a theory on such primary beliefs and judgments is grounding it on easily questionable bases, for one may argue that these are the outcome of social and historical accident, of partiality, or

superstitions. Moreover, it is not impossible to assume that people could hold two set of moral beliefs simultaneously, each of which prescribes different and incompatible responses or moral actions. In such a situation, it would be difficult to see how the method of reflective equilibrium could circumscribe acceptable moral beliefs.

As an example, consider the conclusion that Peter Singer draws in "Famine, Affluence and Morality,"[68] that proximity makes no moral significance on the matter of our moral obligation to promote moral good and to prevent greater moral evils. Singer argues that this conclusion follows from the following premises underscored by the moral beliefs or principles we hold about suffering and death:

> **P1**: Suffering and death from lack of food, shelter, and medical care are bad or great moral evils.
> **P2**: If it is in our power to prevent something bad from happening, i.e. great moral evils without sacrificing anything morally significant (or of equal moral significance), we ought, morally, to do so.
> **P3**: We could give up many of the things that we spend our money on now without morally significant loss.
> **Conclusion**: Therefore, we ought to be contributing some of our money to causes that prevent great moral evils, i.e. suffering and death from lack of food, shelter, and medical, etc.

If the three premises reflect some beliefs that people hold about suffering and death, then Singer's conclusion that we ought to be contributing some of our money to causes that prevent great moral evils and that proximity makes no moral significance on the matter of our moral obligation to prevent greater moral evils

[68] Singer, "Famine, Affluence, and Morality," in *Ethics in Practice: An Anthology*, Hugh Lafollette (ed.), Oxford, Blackwell, 1997, pp.586, 587.

seems persuasive. The implication of this is that no one is morally justified in discriminating against those in need of assistance on the basis of distance. But this seems to conflict with two moral beliefs people hold. The first is the belief about desert and entitlements, i.e. the belief that people are entitled to keep or decide how their property or earnings are expended. The second is the belief about special relationships, i.e. the belief that people owe duties to those to whom they are related and not to strangers, namely, that the duties (to help) to ones' people are special and more extensive and should take priority over duties to other people.

Here we have two sets of beliefs that define a moral situation and yet prescribe two different courses of action. On the one hand, we have a belief that requires that assistance be provided for everyone including strangers and that it is immoral to discriminate on the basis of distance against anyone who needs help. On the other hand, we have the belief about entitlement and the belief that being biased towards the needs and situation of family members, friends, etc is not immoral. Should we employ the method of reflective equilibrium to calibrate acceptable moral beliefs and the principles of justice, as Rawls does in JaF, how should we resolve this conflict or dilemma?

Rawls' tack would perhaps be to deny the assumption that both beliefs are correct. It is not clear how he would frame such denial. In any case, given his constructivist view of moral principles, this tack does not effectively render meaningless the attack on the method of reflective equilibrium. For Rawls, as well as Kant, fundamental moral principles, are not foundational but are constructed from conceptions of the person and of practical reason. And because they are not

foundational, it would be difficult to establish which between two sets of beliefs each of which prescribes different and incompatible moral actions is correct.

Moreover, it might be the case that decision-makers situated in a position of free and equal powers would choose Rawls' principles of justice, but that does not tell us if the elements leading to them reflect relatively uncontroversial fundamental moral beliefs that we hold about the nature of persons and about the role of morality or justice in society? In ToJ, Rawls says, "men agree to share one another's fate."[69] That is, the principle of reciprocity requires that the interests of others are respected in ways that lead to mutual benefits. Since Rawls takes the veil of ignorance to be necessary in achieving the principle of reciprocity and agreement between agents, we can ask, "How do the veil of ignorance and the principle of reciprocity 'reflect a moral belief that we are all supposed to share'"?[70]

Rawls' answer is that one's possession of natural and social assets is "arbitrary from a moral point of view." Inequalities in natural endowments are social evils bearing on the justice of a society. Because they are inequalities of chance in life, people should not suffer or benefit from differences between them for which they are not responsible. Rawls reasoning is that the view that no one deserves to benefit from the accidents of birth reflects uncontroversial fundamental moral beliefs about the role of morality or justice in society, and that this belief is modeled by the veil of ignorance and the principle of reciprocity.

If it is the case that the principles reflect uncontroversial fundamental moral beliefs, then agents have a reason for accepting them since they respect themselves

[69] Rawls, ToJ, Cambridge, MA: Harvard University Press, 1971, p 102. This sentence is eliminated in the 1999 revised edition.
[70] Wolff, p.170.

and others as persons when they accept and act on the principles. There is a further reason for agents to accept and comply with the principles. This is the assumption that agents can be relied upon to comply with principles that they chose in the original position. Recall that Rawls takes the capacities for a sense of justice and a conception of the good as the two moral powers that define the conception of full autonomy and moral persons. The capacity for a sense of justice ensures that agents can be relied upon to comply with principles that they chose in the original position.

The idea of the capacity for a sense of justice is that it would be purely irrational for an agent to take risks, wrongly assuming that if the situation turns out badly, she can either violate the terms of the contract and latter recoup or renegotiate them. This will also apply to agents who find themselves with fewer benefits in virtue of engaging in cooperative ventures or those tempted to appropriate benefits without the corresponding costs. Since agents know that the contract is made in good faith, they do not seek to annul it or have it revoked just because they end up badly when the veil of ignorance is removed.

But is Rawls right? Notwithstanding the forcefulness of Rawls' argument, many people do not share the view that the principles reflect uncontroversial fundamental moral beliefs. For example, many do not share the view that natural assets are "common assets" from which all members of society should gain a benefit. There are those who do not see anything unfair about people benefiting differentially from the employment of their natural endowment. The view that people never deserve to benefit from using their talents and abilities is

questionable. In particular, if someone has worked hard to develop and hone a talent or skill, which they then use to good effect, the moral belief seems to be that they deserve to reap some reward. If this is right, in what way then would the capacity for a sense of justice and the constraint of finality provoke agents into accepting the principles when doing so requires that they ignore the history of legitimate acquisitions and the efforts that they have expended in honing and developing their various talents.

Rawls' view that natural assets are arbitrary and hence common assets and that this reflects relatively uncontroversial moral beliefs commonly held is problematic. As noted above, it conflicts with moral beliefs many people hold about the relationship between talents and desert. In the face of such conflict, it seems indeed correct to say that Rawls has not successfully justified the second condition: the requirement that the elements of the contract and principles match our considered moral judgments about justice in reflective equilibrium. In the absence of such justification, the argument that agents would accept the principles seems implausible.

Moreover, Rawls' discussion of the distribution of natural endowments seems to miss the point about the real status of their emergence. I quote Gauthier at length:

> Rawls talks of the 'distribution of natural talents' and the natural lottery. He falls prey to the dangers that lurk in this talk. There is no natural lottery; our talents are not meted out to us from a pool fixed to guarantee winners and loser. And if there is a distribution there is no distributor—unless we assume a theistic base foreign to Rawls' argument. If there

> were a distributor of natural assets, or if the distribution of factor endowments resulted from a social choice, then we might reasonably suppose that in so far as possible shares should be equal, and that a larger than equal share could be justified only as a necessary means to everyone's benefits. But this would be to view persons as creatures of a distributor—a God or a non-instrumental Society—and not as rational and individual actors. In agreeing with Rawls that society is a cooperative venture for mutual advantage, we must disagree with his view that natural talents are to be considered a common asset.[71]

Rawls' attempted justification of the principles of justice depends on the principles being chosen from the original position and on the original position being drawn up in such a way that whatever comes out will be fair and just. But his failure to (a) argue convincingly for the principle of reciprocity and the veil of ignorance and (b) demonstrate that they reflect an uncontroversial moral belief or judgment that people ought to share because natural endowments are common assets leaves the justification of his two principles of justice problematic. Consequently, JaF seems to fall short of what we need to account for social or distributive justice.

If we accept all the assumptions Rawls makes in JaF: the original position, reflective equilibrium, the maximin rule, indeterminate persons—possessing a capacity for a sense of justice—then not only will Rawls' principles of justice be chosen, but his version of contractualism would constitute a persuasive and fruitful approach both to distributive social justice and to the great contract tradition. However, this is only if we accept all the assumptions in JaF. But we have no

[71] Gauthier, MbA, pp. 220-221.

compelling reason to accept them. Gauthier did not, and it is to his account of moral contractarianism that I turn to in the next chapter and the reminder of this dissertation.

Chapter Three

David Gauthier's Moral Contractarianism—Rationality and Rational Constraints of *Morals by Agreement*

Introduction

As the name clearly indicates, Gauthier's moral contractarianism is a contract-base view of morality, i.e. it is a theory of rational morality. Moral principles, according to Mb(CM)A, are rational constraints on the pursuit of individual self-interest. The fundamental arguments of this brand of contractarianism hang on two related claims: first, that we are primarily self-interested utility maximizers and second, that a rational estimation of the best strategies for maximizing our self-interest will lead us to accept side-constraints on utility-maximization that amount to a rational replacement for traditional morality.

Gauthier appeals to rational choice theory to contextualize and make explicit both of the above claims. Recall that Rawls, too, appeals to rational choice theory. However, while Gauthier uses the resources of that theory to demonstrate how rational self-interest motivates one to act morally, Rawls uses that theory to show how the pursuit of one's interest leads one to show concern for all. Furthermore, whereas Rawls employs the theory of rational choice to ground the need to respect persons as ends in themselves, Gauthier employs that theory to ground the reason to be moral.

In addition to their differing use of the theory of rational choice, two other differences stand out clearly in their social contract accounts. Firstly, the theories differ regarding the sort of agents that make up the contract. For Gauthier,

agreement takes place among determinate individuals who bargain in full view of their preferences, capacities, and circumstances, while for Rawls, the individuals who agree on his two principles of justice possess the capacity for a sense of justice and are hidden behind a veil of ignorance, which screens out their preferences, capacities, and circumstances. The veil of ignorance separates the real self or nature from its aims or ends. The self, which is stripped of all contingencies and individuality is prior to the ends which are affirmed by it. What reveals our nature, according to Rawls, is not our aims but "the principles that we would acknowledge to govern the background conditions under which these aims are to be formed and the manner in which they are to be pursued."[72] Gauthier agrees with Rawls that our nature is prior to the ends, which are affirmed by it. However, unlike Rawls he claims that our ends are primarily revealed by our nature. We do not set aside our preferences, rather our preferences serve as motivation for the moral principles we agree to.

Secondly, the subject matter of their social contract differs. Whereas Rawls wants an agreement on principles of distributive social justice that are intended to govern the basic structure of society, Gauthier wants an agreement on moral principles to govern individual relationships. To this extent, it is appropriate to consider Justice as Fairness as a political project, and morals by constrained maximization agreement a moral project.

Gauthier's moral contractarianism has several attractive features. It is theoretically more fundamental than JaF. Speaking of its theoretical rigor

[72] Rawls, ToJ, p.491.

Braybrooke hails it as the best and most promising in the great social contract tradition. He writes:

> Social contract theory flies higher and more expertly in *Morals by Agreement* than ever before. But it does not fly alone. Gauthier has carried to the same height of sophistication the project, often mooted by philosophers past and present, of deducing morality from rationality. The theory of the social contract is the most promising vehicle for the deduction project, bringing together the themes, all crucial to ethical theory, of consent, mutual benefit, and cooperation. No one previously has come anywhere as near as Gauthier to carrying the project through with perfect precision and rigor, incidentally-a third triumph-making better combined use than utilitarianism itself of the notions of utility and optimization.[73]

One cannot speak of the theoretical rigor and sophistication of Mb(CM)A without reference to the science of rational choice theory that it is wedded to. That theory treats practical reasons as 'strictly instrumental' to the satisfaction of preferences about outcomes. In contrast to Rawls who applies the principles of rational choice to hypothetical choice situations or idealized conditions, Gauthier applies the principles to real-life or determinate choice situations. That is, whereas Rawls applies the principles to agents who in the original position begin with considered moral judgments that arise from a sense of justice, Gauthier applies them to agents who possess determinate preferences and who are interested in choosing moral principles that advance those preferences.

[73] Braybrooke, "Social Contract Theory's Fanciest Flight", *Ethics: An International Journal of Social, Political, and Legal Philosophy*, vol. 97, no 4, July 1987, p. 751.

Gauthier's commitment to rational choice theory is deeper and more robust than Rawls.' Recall that, for Rawls, in justifying a particular description of the original position, the idea is to see if the principles that are chosen by us provide guidance where guidance is needed, and "match our moral judgments or considered convictions of justice or extend them in an acceptable way."[74] These judgments, which may be judgments about general or specific moral principles or cases, may conflict in some way, and when they do, we proceed by adjusting our various beliefs and ideas until they are stable or in equilibrium. Moral beliefs that are stable or in ideal reflective equilibrium provide consistent practical guidance and they, according to Rawls, describe the fundamental principles of our sense of justice. But for Gauthier, the moral principles that we choose do not require any balancing with our considered moral judgments or convictions in reflective equilibrium, rather we choose the moral principles based on our preferences or overarching interest of utility-maximization, constrained of course by the requirement of rationality. These principles, which sufficiently guide social interaction and practices offer us a compelling reason to constrain our pursuit of strict maximization of individual utility, a quantity that is associated with preference.

Given that Gauthier allows his agents to have knowledge of their preferences—and to permit these to influence the choice of principles—he and not Rawls shows more commitment to the theory of rational choice, which takes individual preference as basic. Rawls's agents do not reason about utility; they reason about social primary goods. Gauthier's agents reason not about social primary goods but about utility, which explains preferences. Rational choice theory

[74] Ibid, pp.17, 18.

identifies rationality with the maximization of utility, which it defines not as a measure of 'real' preferences related to actual circumstances, but as a measure of coherent considered preference about outcomes. Grounding Mb(CM)A on rational choice theory, Gauthier takes moral constraints as arising from the preferences of bargainers, for the purpose of maximizing expected utility.[75]

Because Gauthier applies the principles of rational choice to real-life choice situations and determinate agents, he, unlike Rawls, is able to demonstrate how to proceed if one must generate strictly rational principles without introducing prior moral presuppositions.[76] In attempting to derive morality from rationality, Gauthier's strategy is to develop a set of constraints mandated by practical rationality and to identify these constraints as moral principles. He does this by first linking morality to reason, then reason to practical reason and practical reason to interest, i.e. individual utility. Gauthier's argument is that if morality is to convince

[75] The deduction project, i.e. deducing morality from rationality, according to Braybrooke, marks one significant improvement of Mb(CM)A over JaF. The latter, he says, appeals to question-begging assumptions, which the former jettisons. One of those question-begging assumptions, according to Braybrooke, is 'the formal constraints of the concept of right', which he employs to circumscribe the range of principles considered in the original position. Unlike Gauthier, Rawls chooses to pursue a different, less radical deduction project. His deduction project takes the choice of principles to proceed from rationality constrained by prior moral considerations or judgments. The point, according to Braybrooke, "is that Gauthier does not assume any of Rawls's constraints, either as regards bargaining or in the original position argument. Gauthier's agents simply are to maximize their utilities in respect to cooperation (or have their utilities maximized for them, by the Archimedean choice), and from that aim Gauthier deduces the attraction of an impartial scheme of constraints and benefits that is just." Braybrooke, "Social Contract Theory's Fanciest Flight," in Ethics, p.755.

[76] Rawls' reflective equilibrium, veil of ignorance, the capacity for a sense of justice, moral intuition or considered moral beliefs, are some examples of prior moral assumptions that Gauthier jettisoned in Mb(CM)A. There is an ongoing debate in the literature as to whether Gauthier really succeeds in deriving morality from rationality, i.e. whether his starting premises are actually non-moral or morally neutral. See Holly Smith's argument that a close examination of the Mb(CM)A reveals that either its moral principles are not genuine moral principles or that the principles of rationally it appeals to are morally laden; and Jan Narveson's discussion of the sense in which Gauthier's appeal to the 'equal rationality of bargainers' might be understood as not morally neutral. Smith, "Deriving Morality from Rationality" (pp.229-253) and Narveson, "Gauthier on Distributive Justice and the Natural Baseline" (pp.127-148) in *Contractarianism and Rational Choice: Essays on David Gauthier's* ***Morals by Agreement***, Peter Vallentyne (ed.), New York, Cambridge University Press.

rational agents, or speak to their affective capacities, or otherwise solve the problem of rational compliance, or indeed completely move them to actions it must be grounded on assumptions that are not only widely acceptable but are themselves not moral; morality cannot assume what it sets out to prove.

Contractarianism, Gauthier argues, offers the only plausible resolution to the foundational crisis facing morality in our modern world. The crisis of morality concerns the justifiability of impartial prescriptive or moral principles that are in some sense universal in scope, i.e. "a lack of fit between what morality presupposes — objective values that help explain our behavior and the psychological states — desires and beliefs — that, given our present world view, actually provide the best explanation."[77] This crisis has been recognized by a number of philosophers[78] and is captured by Friedrich Nietzsche's remark in *On the Genealogy of Morals*, "As the will to truth thus gains consciousness — there can be no doubt of that — morality will gradually *perish* now."[79]

Nietzsche's remark is best summarized as follows. There are two worldviews: our present view of the world and an old view of the world. Moral language presupposes the old worldview—a view of the world as purposively

---

[77] Gauthier, "Why Contractarianism?" in *Contractarianism and Rational Choice*, Ibid, p.16.

[78] (1) "There are no objective values…. [However] the main tradition of European moral philosophy includes the contrary claim" – John L. Mackie, *Ethics: Inventing Right and Wrong*, Harmondsworth, Penguin, 1977, pp.15, 30; (2) "Moral hypotheses do not help explain why people observe what they observe. So ethics is problematic and nihilism must be taken seriously" – Gilbert Harman, *The Nature of Morality*, New York, Oxford University Press, 1977, p.11; (3) "The hypothesis which I wish to advance is that in the actual world which we inhabit the language of morality is in… [a] state of grave disorder…we have – very largely, if not entirely – lost our comprehension, both theoretical and practical, of morality" – Alasdair MacIntyre, *After Virtue*, Notre Dame, IN, University of Notre Dame Press, 1981, p.2; (4) "The resources of most modern moral philosophy are not well adjusted to the modern world" – Bernard Williams, *Ethics and the Limits of Philosophy*, Cambridge, MA: Harvard University Press, 1985, p.197.

[79] Friedrich Nietzsche, *On the Genealogy of Morals*, 3rd edition, translated by Walter Kaufman and R.J. Hollingdale, New York, Random House, 1967, section 27, p.161, emphasis in original.

ordered. Our present worldview and not the old one best explain our psychological states. Given this fact and given that we have abandoned the old worldview, the moral language that goes with it must equally be abandoned, and abandoning the old worldview along with the moral language that presupposes this worldview essentially amounts to the perishing of morality.

Gauthier's moral contractarianism challenges this outlook by attempting to "allay the fear, or suspicion, or hope, that without a foundation in objective value or objective reason, in sympathy or sociality, the moral enterprise must fail."[80] In resolving this crisis, Mb(CM)A grounds human behavior, choices, and actions in coherent considered preferences of individuals.[81] By grounding choices on considered preferences, Gauthier jettisons accounts of value that are objective. Gauthier takes value to be subjective.

There is an important reason why he thinks we must embrace a subjective account of value. A subjective account of value, he claims, explains better our behavior and our psychological states. To think of value as objective is to regard it as existing independently of the preferences of agents and as providing a standard to govern preferences. Gauthier's subjective account of value identifies value with utility and takes it as a measure rather than the norm for rational preference. On this account of subjective value and utility, we take an action or disposition—for, as we shall see later Gauthier replaces actions with dispositions—if and only if that action or disposition maximizes expected utility.

---

[80] Gauthier, "Moral Artifice: A Reply by Gauthier" in *Canadian Journal of Philosophy*, vol. 18, 1988, p.385.
[81] Gauthier, "Why Contractarianism?" in *Contractarianism and Rational Choice*, pp. 15-25.

It is important to note that a subjective account of value does not imply that whatsoever an agent maximizes insofar as it is indexed to utility must be the measure of preferences. Satisfied preferences must meet the objective conditions of preferences, i.e. preferences must be objectively valuable. Preferences are objectively valuable when utility measures not just revealed preferences but also attitudinal or 'expressed preferences.'[82] To state differently the point about preferences and utility we can think of it in terms of subjective and objective determination of reasons for acting. In acting, individuals have reasons motivated primarily by their beliefs, desires, and emotions. Although these reasons for acting will largely confirm and be confirmed by what other agents consider as reasons what is decisive is the individual's own determination.[83] The subjective determination explains the individuality of agents and the objective determination we might say explains the objective condition of reasons.

Moral principles are rational constraints on individual behavior. The reason why they so constrain is because they arise from agreements made between rational agents for the purpose of advancing self-interest. But why should we accept constraints on our behavior? The idea is that as utility-seeking agents, we desire or aim to maximize the possible outcomes of actions (or dispositions) and the reason we accept or ought to accept moral or rational constraints on our behavior is that those constraints facilitate cooperation. So for anyone who asks the question: "why must I accept the constraints of morality?" The answer, according to Gauthier, is

[82] MbA, p. 28.
[83] Gauthier, "Moral Artifice: A Reply by Gauthier", p.388.

clear and simple: "the constraints make possible the conditions that are necessary for pursing your rational self-interests, i.e. for achieving your desires or aims."

If the outcomes that moral principles make possible are those that we desire and value, then it is rational for us to embrace the constraints that they impose. If we embrace moral constraints for the outcomes they make possible, then it is right to say that the constraints are self-imposed, i.e. they arise from our considered coherent preferences and not imposed by some eternal conceptual relations that hold true independently of us. Because Gauthier's strategy in Mb(CM)A ties moral constraints with preferences he can boldly declare that (moral) contractarianism provides the most plausible resolution to the crisis of morality in our modern world. Moral contractarianism, according to Gauthier, does more than resolve the crisis of morality. It provides a perceptive strategy for resolving the general problem of rational compliance.

Some of the dominant moral theorists—the rationalists, who side with Kant and construe moral constraints as natural, i.e. drawn from our reasoning faculties, and the naturalists, who side with Hume and take morality to be constructed from shared sentiments—fail, according to Gauthier, to address both the crisis of morality and the problem of rational compliance.[84] These theories fail the demand of the rational skeptic because they do not speak to the interests of the skeptic. It is one thing to say that moral terms and obligations are essentially indexed to reasoning faculties or benevolent sentiments and it is another thing to show how they provide sufficient motivation or justification for their acceptance. So while

[84] See Gauthier, "Why Contractarianism?" in *Contractarianism and Rational Choice*, pp. 18-19. In section 3 of chapter four, I will discuss Hume's naturalistic moral sentiments as it relates to Gauthier's moral contractarianism and the issue of secession.

one might claim that reason and sentiment provide the templates for morality, one cannot guarantee that people would embrace the constraints that arise from them. For if agents do not identify with these faculties the constraints that emerge from them would be useless if not empty.

In order to address the general problem of rational compliance Gauthier, in MbA and elsewhere, situates morality within the realm of the artificial and takes the view that moral concepts arise "for rational persons, only within a framework of agreed constraints."[85] Gauthier's moral contractarianism is a sort of conventionalism and is understood in terms of what agents would hypothetically agree to subject, of course, to some initial constraints, which are themselves not the product of agreement but the precondition of rational agreement. The initial constraints are taken as the condition for the possibility of any agreement, and the rationality of agents is defined to the extent they agree to such constraints.

Agreement is the basis of moral constraints and these constraints move us to action when agreement starts from a morally constrained initial position that does not appeal to moral concepts or reasons. This is what Gauthier means when he says that "morality we shall argue can be generated as a rational constraint from non-moral premises of rational choice."[86] This method of generating moral constraints marks a fundamental difference between Gauthier and some other contract theorists, especially Hobbes, regarding what it is that supports constraints. Whereas moral constraints, according to Gauthier, are self-supporting in the sense that it is the internal mechanism of the constraints themselves that present them as

[85] MbA, p.85.
[86] Ibid, p.4.

rational to us, for Hobbes, it is the external mechanism of a coercive political system that supports them. This view of self-supporting moral constraints is a fruitful one for Gauthier since it provides him the basis to argue the claim that moral contractarianism provides the only possible resolution to the crisis of morality. Thus, according to Gauthier, if moral constraints appeal only to rationality—our considered coherent preferences—and not anything outside of it, then contrary to Nietzsche's prophesy, "as the will to truth thus gains consciousness" morality will *not* perish.

3.1 The Demand of Morality and the Three Principal Sub-theories in Mb(CM)A

In chapter one, I examined among other things what it means for social contract theory to ground moral and political norms or principles on agreement. I discussed the process by which such principles arise and the role they play in society. I now want to examine how by "adding the rigor of rational choice"[87] to the hypothetical contract Gauthier teases out moral principles. I hope to be able to show at the end of my discussion how Mb(CM)A's rigorous appeal to the theory of rational choice leads to moral principles that weld together separate, unattached, disparate, and mutually unconcerned individuals whose interests frequently conflict, and how this approach fares firstly, under the test of the general problem of rational compliance, and secondly, under the test of application of the problem of secession.

If we think of the raison d'être of morality as providing reasons that override the reason of self-interest in those cases wherein it is detrimental to all for

[87] Ibid, p.10.

everyone to follow his or her self-interest, then moral constraints can be viewed as necessary to the pursuit of self-interest. Morality, we suppose can be either 'grandiose' or 'adequate.' It is grandiose if it provides answers to every moral question and leads to a perfectly moral society. It is adequate if it helps justify human self-regarding interests. Given our knowledge of humans and society, and given the foundational crisis of morality, there is no reason to suppose that any moral theory, for that matter, can meet the grandiose demand. But can a moral theory in general and moral contractarianism in particular meet the adequate demand of morality?

A morality is adequate if it passes three related tests: the compliance test, the contractarian test and the efficiency test. The compliance test is a demand for the demonstration of the rationality of accepting moral constraints. The contractarian test requires that every person must voluntarily accept the constraints. And the efficiency test is a demand that the outcome of agreement be optimal. Note that what the triple tests of the adequate demand of morality require is that morality *must* serve our self-regarding interests. Morality serves an individual's interests when, for that individual, the marginal benefits that the scheme of cooperation that morality engenders is not less than the marginal costs that such individual expends in support of that scheme. The adequate demand of morality thus embodies the general problem of rational compliance. Does Mb(CM)A meet the three demands of morality? Gauthier says this about the constraints of moral contractarianism:[88]

[88] Gauthier, "Moral Artifice: A Reply by Gauthier", pp.288, 389.

(a) Its constraints are not externally imposed, but based on what rational agents reflectively accept, for morality cannot be an impostor.
(b) Mb(CM)A constraints are voluntarily accepted, for morality cannot be coercive as a system of domination.
(c) Its constraints lead to optimal outcomes, hence, overcomes the structural problem of suboptimality in natural interactions.

Whereas *c* constitutes part of the broad demand of mutual advantage and the requirement that the costs imposed on any individual be proportional to the benefits that the individual receives, *a* and *b* represent the underlying idea behind the thesis of individualism. If it turns out that Mb(CM)A as Gauthier claims satisfies the triple tests of the adequate demand of morality, then it is indeed clear that he has fashioned out a perspective of rational morality that not only withstands Nietzsche's challenge but resolves as well the general problem of rational compliance.[89]

Does Mb(CM)A satisfy the three tests of the adequate demand of morality? I now want to examine this question within the general rubric of Mb(CM)A's principal components or sub-theories. The first sub-theory is the principle of rational agreement or a theory of rational bargaining, which Gauthier identifies as the Minimal Relative Concession principle. The second sub-theory is the principle of constraints or a theory of rational compliance, otherwise called Constrained Maximization. And the third, the principle of natural property rights or the theory of the appropriate natural baseline for the social contract, which is simply known as the Proviso. Gauthier employs the first sub-theory to solve the bargaining problem, the second to address the general problem of rational compliance, and the third to

[89] Ibid, pp.288, 389.

resolve the contract problem. Whereas the second sub-theory directly addresses the general problem of rational compliance, the first and third sub-theories contribute to illuminating it. I begin with the first sub-theory, the Minimax Relative Concession principle.

3.1.1 Minimax Relative Concession and the Bargaining Problem

Gauthier's discussion of Minimax Relative Concession (MRC, for short) is in Chapter V of MbA. MRC, he claims, is a unique bargaining principle that captures the idea of justice and impartiality in bargaining situations. Gauthier begins by defining justice and then proceeds by developing a bargaining principle that he says is compatible with such a conception of justice. Justice, he says, is "the disposition not to take advantage of one's fellows, not to seek free goods or to impose uncompensated cost, provided that one supposes others similarly disposed."[90] If justice prohibits taking advantage of one's fellows and imposing uncompensated cost on others, then a bargaining theory is compatible with this conception of justice if it measures up to an agent's expectation of the benefits from cooperation.

To this extent, a bargaining principle must satisfy two conditions. The first is the 'condition of improvement.' This condition states that each bargainer must get some portion of the cooperative surplus, i.e. a bargainer *must* leave the bargain with more than he or she had prior to it or more than he or she brought to the bargaining table. Relating marginal benefits to marginal costs, we say that for every person, marginal benefits must not be less than marginal costs. The second is

[90] MbA, p.113.

the 'condition of parity.' According to this condition, each bargainer must "take from the bargain the expectation of some utility at least equal to what she would expect from non-cooperative interactions."[91] Both conditions express in different form the general idea of mutual advantage. If a bargaining theory meets these conditions, the requirement of justice, Gauthier argues, is satisfied. MRC, he claims, meets the double conditions.

Gauthier defines MRC as the measure of a person's stake in a bargain situation. That is, the difference between the least each person would accept in place of no agreement and the most each person receives in place of being excluded by others from the agreement.[92] Gauthier defines the concept of relative concession in this way:

> If the initial bargaining position affords some person a utility u^*, and he claims an outcome affording him a utility $u\#$, then if he concedes an outcome affording him a utility u, then the absolute magnitude of his concession is $(u\# - u)$, of complete concession $(u\# - u^*)$, and so the relative magnitude of his concession is $[u\# - u)/ u\# - u^*)]$.[93]

The general bargaining problem is the problem of what principle it is rational to adopt in market failure, i.e. situations of individual and collective actions or in the production of public goods.[94] The specific problem that a bargaining theory attempts to solve is the problem of choosing among a number of

---

[91] Ibid, p.133.
[92] Ibid, p.14.
[93] MbA, p.136.
[94] Gauthier states the problem of market failure in this way, "Where the invisible hand fails to direct each person, mindful only of her own gain, to promote the benefit of all, cooperation provides the visible hand. … We begin our examination of cooperation as the rational response to market failure," Ibid, p.113.

possible but mutually incompatible distributions of the cooperative surplus. The cooperative surplus is the sum of benefits made possible by individual contributions or cooperation. Given that a person's participation in cooperative activities is presented in the form of opportunity cost, what that person gets from the cooperative surplus is determined by what she could or ought to get in her next-best available option—the outcome by not participating in cooperation.

As an example, consider Abel and Mabel,[95] both of whom can make 5% per annum on their money in a savings account. However, if they pool their money and invest it in a money market account they will make 10% annually. Let us suppose that at least $700 is needed for the investment. Suppose also that Abel has $400 and Mabel $600. Suppose finally, that all of what Abel and Mabel have were invested. If all $1000 were invested, the cooperative surplus would be $50.[96] What then is the bargaining problem? The bargaining problem that confronts both of them is how to divvy up the cooperative surplus, that is, who gets what or how to share the $50.

First, we identify the cooperative surplus. Next, we characterize the claim on the surplus that each person makes. Note that each person's claim and concession are comparable to the claims and concessions of others. Since Gauthier

---

[95] This example draws substantially on Jean Hampton's adaptation of Gauthier's example in "Equalizing Concessions in the Pursuit of Justice: A Discussion of Gauthier's Bargaining Solution" in *Contractarianism and Rational Choice*, pp.148-161.

[96] Money in saving account (yields 5%):
For Abel = 5% of $400 yields $20
For Mabel = 5% of $600 yields $30
Money market investment (yields 10%):
For Abel = 10% of $400 yields $40
For Mabel = 10% of $600 yields $60
The cooperative surplus is therefore the difference between yields in the money in savings account for Abel and Mabel and yields in money market investment, which is $50 ($20 + $30) – $100 ($40 + $60).

believes that a cardinal measure of intrapersonal and interpersonal utility is possible and that utilities are linear with monetary values, he thinks comparative claims and concessions can be evaluated. If we assume that this is the case, then it is possible for both Abel and Mabel to evaluate the claim and concession each person makes.

The bargaining process reminds us of the hagglers in a typical African market. It begins when each party advances an initial claim for some portion of the cooperative surplus or puts forward a claim for a bargainer's rate of return. If these claims are compatible, then agreement is reached and the bargaining process ends. But as more likely, the claims would be incompatible, hence a second stage in which each party offers a concession to the others by withdrawing some portion of her original claim and proposing an alternative outcome. Gauthier argues that the bargainers, we suppose, are equal and rational. Given this, the maximum concessions, i.e. the greatest proportion of their original claims that they give up "must be minimized."[97] They must not concede so as to leave them with higher marginal costs relative to the marginal benefits they receive. The process of making concession continues until a set of mutually compatible claims is reached.

How much is it rational for each party to claim? Each person's claim, Gauthier argues, is bounded by the overall cooperative surplus and by the portion of the surplus that is possible for that person to receive, determined by the extent of her participation in cooperative interactions. Since we suppose that cooperation is better for everyone than noncooperation, each person must concede as much as is

[97] Gauthier, "Justice as Social Choices", in *Social Contract Theory*, (Michael Lessnoff (ed.), Oxford, Basil Blackwell, 1990, p.204.

necessary to make the former possible. Each person must not concede so little as to be excluded from the benefits of the cooperative surplus or to deadlock agreement.

The point about concessions and benefits in a bargain is well illustrated by the example in chapter one of employing either a private will policy or a general will policy. If each employee acts on the former, no one gets any of the $1M for workers' bonuses, but if each acts on the latter, each employee individually gets a portion of the $1M available. To get a portion of the $1M is to put a claim for $10,000. To put a claim for $10,000 is to concede some portion to others, and to make a concession to others is to cooperate. To this extent, the MRC rule serves as a guide to agents on how best to maximize their gains as they engage in cooperative activities with others. Hence, it is rational for each agent to aim for a happy medium point. The happy medium point, according to Gauthier, is the point at which each bargainer makes concessions that are (as nearly as possible) proportionate or equal to the concession of other bargainers.

The happy medium point is a midway between two points or extremes; it embodies, so to speak, a satisfactory outcome. Gauthier thinks that the stability of a scheme of cooperation depends on whether each bargainer's happy medium point is met—it is no use having a malcontent or a disgruntled bargainer. Since $50 is the cooperative surplus available to Abel and Mabel, the two points or extremes are $0 and $50, hence, the happy medium point for each is $25.

If we accept the claim that the stability of a scheme of cooperation depends on whether each bargainer's happy medium point is met, then the bargaining process must maneuver the concessions that Abel (or Mabel) makes relative to the

claim and contribution Mabel (or Abel) makes and to the cooperative surplus. Ideally, each person would want the entire cooperative surplus, i.e. each would concede the smallest by insisting on the total gain. However, insisting on the entire $50 would likely deadlock the agreement and make cooperation between them impossible. Again, going back to the private will policy versus general will policy example, putting forward a claim for $50 is equivalent to each employee insisting on taking home the entire $1M available for workers' bonuses. But to insist on taking home all $1M is to claim something ($100M) that is not on offer. For Abel or Mabel to each claim $50 is therefore to claim $100, which is not on offer. A claim of $100 will more than likely deadlock the agreement, and in the absence of agreement there is no cooperation.

The biggest concession would be for Mabel to settle for none of the surplus while Abel goes home with all $50. But it would unreasonable for Mabel to do so since that allows Abel to have a partial freeride. Allowing Abel to go home with $50 means Mabel incurs more costs from the investment than she benefits. We would expect Mabel not to agree to this situation, given that impartiality in market and cooperative interaction requires that each person "has a sufficient reason to consider interaction with his fellows to be impartial only in so far as it affords him a return equal to the services he contributes through the use of his capacities."[98] A situation that affords Mabel none of the cooperative surplus fails both conditions of improvement and parity, which stipulates that for cooperation to be rational and just those who are party to weaving its fabric must benefit from it. Because the

[98] MbA, p.100; See section 3.3 of chapter IV of MbA for Gauthier's detailed discussion of freeriding and the problem it poses for the market and cooperation.

MRC principle requires that no one be taken advantage of, it requires that each person make concessions that are as nearly as possible equivalent to the concessions of others. Since the MRC rule weighs the concessions of a bargainer relative to those of others, the happy medium point for Abel and Mabel would be around $25, i.e. ½ and ½.[99]

In order to prevent others from taking advantage of them and from maximizing their utilities at their expense, bargainers, according to Gauthier, must stay well within the happy medium point. Since we suppose that each bargainer rationally weighs her[100] concessions relative to the claims and concessions of other bargainers and since we suppose too that a bargainer's utility profile is not maximized if her happy medium point is significantly comprised, each bargainer, we assume, would generally ensure that her concessions to the others are as small as possible. This is especially the case if we suppose that the initial or natural baseline for bargain is determined, as Gauthier says, by what each agent brings from the pre-contract stage, i.e. her initial factor endowment. Ignoring the input of

---

[99] **For Abel**
u# = 470 [400 + 20 (the amount he could make on his own) + 50 (the entire cooperative surplus)]
u = 445 [400 + 45 (the amount that Abel concedes)]
u* = 420 [400 + 20 (the amount Abel could make on his own)]
Therefore
u# – u = 470 – 445 = 25 = 1 or 0.5
u# – u* 470 – 420 50 2

For Mabel
u# = 680 [600 + 30 (the amount she could make on her own) + 50 (the entire cooperative surplus)]
u = 655 [600 + 55 (the amount that Mabel would get if Abel concedes $45)]
u* = 630 [600 + 30 (the amount Mabel could make on her own)]
Therefore
u# – u = 680 – 655 = 25 = 1 or 0.5
*u# – u** 680 – 630 50 2

[100] My use of the pronoun her here is for convenience and has no sexist overtone.

agreement, we can therefore, say that a bargainer's rate of return is directly proportional or bears some relationship to (1) that bargainer's initial factor endowment, namely, what that bargainer brings to the bargaining table or that bargainer's contribution to the cooperative surplus and (2) the portion of the cooperative surplus to which other bargainers lay claim.

3.1.1.1 Is MRC a Unique Distributive Principle?

One can criticize the MRC principle on the ground that it is not a unique distributive principle in the bargain stage of the contract.[101] The criticism goes this way. The division of the cooperative surplus made possible by the MRC principle, according to Gauthier, is in accordance with what rational bargainers would accept as impartial and fair because it "expresses their equal rationality."[102] This to be sure might be the case, but it can be shown that there are other principles that distribute the cooperative surplus that would be accepted by bargainers as impartial and fair. Therefore, the MRC principle is not a unique principle that would be chosen by rational bargainers. Jan Narveson also raises a similar point when he argues that "the question of why we should be settling in the middle rather than somewhere else [as the MRC principle stipulates] as a supposed matter of "reason"... [is] acute. The appeal to reason seems inappropriate"[103] given "broadly Lockean rights in ourselves and in external item of property."[104]

---

[101] For variants of this criticism see for example Russell Hardin's "Bargaining for Justice," in *Social Philosophy and Policy*, vol. 5, 1988, pp.63-74, and Jean Hampton's "Equalizing Concessions in the Pursuit of Justice: A Discussion of Gauthier's Bargaining Solution." pp.149-161.
[102] MbA, p.143.
[103] Narveson, "Gauthier on Distributive Justice and the Natural Baseline" in *Contractarianism and Rational Choice*, p.132.
[104] Ibid, p.136.

Several candidates for the distributive principle of justice have been proposed. Among these are the Equal Rate of Return Principle or the Principle of Proportionality and the Principle of the Effect of Contributions. Under the Principle of Proportionality (PP), each person receives that portion of the cooperative surplus that is proportional to her contribution.[105] If we apply this principle to the money market investment of Abel and Mabel, Abel would receive $20 for his $400 contribution, while Mabel would receive $30 for her $600 contribution.[106] One reason why this principle seems attractive is that it is congenial with the idea that the amount or level of contribution ought to factor into the principles of reward or payments.

On the other hand, the Principle of the Effect of Contributions (PEC) rewards each person according to the role played by his or her contributions in actually securing the cooperative surplus.[107] In the application of this principle,

---

[105] This is the principle that Hampton endorses.

[106] **For Abel**

u# = 470 [400 + 20 (the amount he could make on his own) + 50 (the entire cooperative surplus)]

u = 440 [400 + 40 (the amount that Abel concedes)]

u* = 420 [400 + 20 (the amount Abel could make on his own)]

Therefore

$\frac{u\# - u}{u\# - u^*} = \frac{470 - 440}{470 - 420} = \frac{30}{50} = \frac{3}{5}$ or 0.6

For Mabel

u# = 680 [600 + 30 (the amount she could make on her own) + 50 (the entire cooperative surplus)]

u = 660 [600 + 60 (the amount that Mabel would get if Abel accepted $40)]

u* = 630 [600 + 30 (the amount Mabel could make on her own)]

Therefore

$\frac{u\# - u}{u\# - u^*} = \frac{680 - 660}{680 - 630} = \frac{20}{50} = \frac{2}{5}$ or 0.4

[107] Narveson does not endorse either PP or PEC. His view is that the MRC principle is "a subordinate and largely dispensable principle", i.e. it "makes sense only against the background of independent rights of the parties concerned: rights to their own person in the way of abilities and other resources, and rights to assorted items of external property." Given what he says above, it seems right to say that he supports a scheme that leaves the distribution of the benefits of

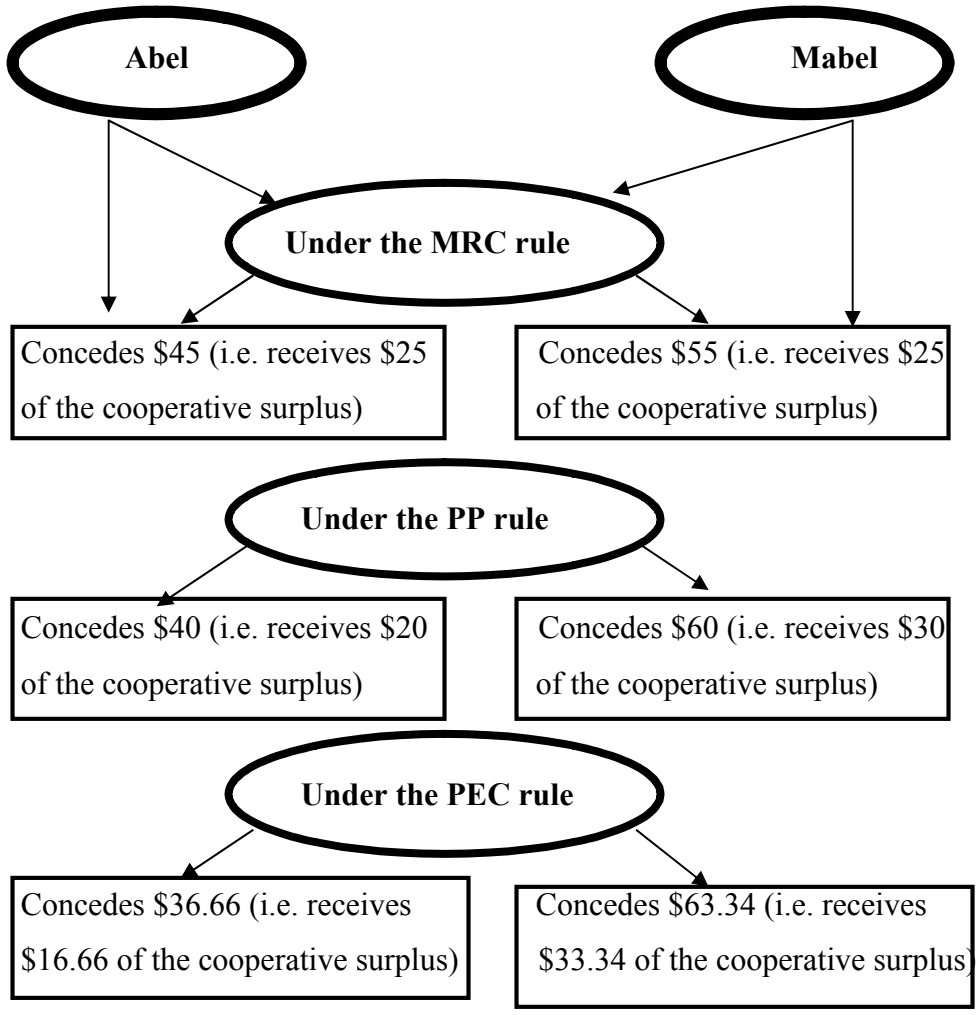
Mabel receives more than she did under PP and Abel receives less than he did under PP. Under PEC, Mabel gets $33.33 or 2/3 and Abel gets $16.66 or 1/3 of the $50 cooperative surplus. Mabel receives $33.33 because her $600 yields 2/3 of the cooperative surplus and Abel receives $16.66 because his $400 yields 1/3 of the cooperative surplus. If we apply all there principles (MRC, PP, and PEC) to the money market investment of Abel and Mabel we get the following, as is shown in figure 3.1.1.1.

Notice that although both principles appeal to contributions in parsing out what each person gets, PEC is slightly different from PP. PP considers only the proportion of contributions among contributors and not the role played by the contributions in producing the cooperative surplus. Stated differently, PP evaluates the contributions in terms of both the proportion of contributions of each party and what each person would have received were he or she not to cooperate. In contrast, PEC evaluates the contributions in terms of the cooperative surplus by correlating the percentage of contribution to the surplus. If we apply all there principles (MRC, PP, and PEC) to the money market investment of Abel and Mabel we get the following, as is shown in figure 3.1.1.1.

---

cooperation or the principle of distribution (of contributions) to bargainers, to what they consider impartial and fair, or to a scheme that allows the dynamics of the contract determine desert. He says, "Gauthier's characterization of MRC as a principle of 'justice' puts it on the same level as, say, the principles of promise keeping, truth telling or fair dealing.... The validity of a promise is surely not due to its approximating an equal division of anything. If the background conditions are properly observed, and everything is on the up-and-up, then promises and contracts are valid by virtue of their form, of the act of agreement itself, and not in virtue of the resulting distributions exemplifying some or other proportions." Narveson, "Gauthier on Distributive Justice and the Natural Baseline" in *Contractarianism and Rational Choice*, p.135. See also, Narveson, *The Libertarian Idea,* Ontario, Broadview Press, 2001, p.196.

Figure 3.1.1.1: **Payoffs for Abel and Mabel under the Three Principles of Distribution**



Where does this leave us? Since all three principles satisfy the requirement of impartiality and fairness, there does not seem to be any compelling reason why one principle has to be chosen over the others. In a situation like this, either the bargain is deadlocked or a principle ends up being chosen that does not command the agreement of all. For example, given the utilities that the three principles afford him, Abel would prefer MRC to PP, which he prefers to PEC. Conversely, Mabel would prefer PEC to PP, which she prefers to MRC. In the absence of a mechanism

to resolve these differences, the bargain would likely be deadlocked. If they cannot agree on a bargaining principle they might as well bid cooperation farewell.

Perhaps both Abel and Mabel can institute a lottery scheme to pick out a particular distributive principle. Suppose they do. And Suppose the scheme picks out the MRC principle is there any reason to consider it a fair principle of distribution? If both of them have agreed to accept the outcome of the lottery scheme, and if we suppose that the scheme picks out the MRC principle, then either might reasonably acquiesce to the distribution brought about by the principle. I say might because the fact that the lottery scheme picks out any principle, in this case the MRC principle it does mean that Mabel might not see it and the distribution the principle engenders unfair. Mabel might consider for instance that the principle is unfair on the ground that it parcels benefits in ways that fail to take into account her contributions.[108] We might agree that heads you get the smallest piece of cake and tails I get the largest piece. It might turn out that you get the smallest piece because the coin turned heads. And we would expect you to settle for the smallest piece given the result of the coin flip, but that does not mean that you might not consider the result unfair, particularly if the cake was paid for by you or if you contributed half the price. If the MRC principle turns out to be Mabel's least favored principle it is so because in addition to it discounting her contributions it generates a scheme of cooperation that leaves her with fewer

[108] There is also the side issue of whether or not the MRC principle does not legitimate parasitism or freeloading. See Hardin's "Bargaining for Justice" for his discussion of how he thinks MRC legitimates parasitism in the areas of distribution that leave out the full range of side payments, pp.67-70. See also Hampton's "Equalizing Concessions in the Pursuit of Justice: A Discussion of Gauthier's Bargaining Solution" in *Contractarianism and Rational Choice*, pp.151-155 for his discussion of Abel's parasitism on Mabel.

benefits. Hence, she might rightly consider it as well as the scheme of cooperation that it generates unfair, and thus reject any bargaining process that incorporates the MRC principle.[109]

Even if Mabel accepts the MRC principle on the ground that she had agreed to the lottery scheme that picked it out it does not mean that she cannot insists that the contact be renegotiated. This line of reasoning is available to Gauthier because he, unlike Rawls, does not assume that decision-makers possess the capacity for a sense of justice. In Mb(CM)A there is no constraint of finality imposed on the bargain and on the moral principles that are chosen. In general, the contract for Gauthier is not a once-in-a-lifetime event; it is not an episode that is not going to be replayed. In particular, the bargain is not a case of for better or for worse for bargainers. That the contract for Gauthier is not a once-in-a-lifetime event means that a bargainer that gets the wrong end of the stick during the bargain is not necessarily doomed to put up with it for the rest of his or her life. There is always room for bargainers to renegotiate the contract.

Note however, that I am assuming that Abel cannot compel Mabel to accept the MRC principle of distribution; the same way that she cannot compel him to accept the principle. To compel her to accept the principle is to violate part of the thesis of individualism according to which moral constraints must not be based on a coercive system of domination but on a system of voluntariness. In saying this, I am not denying the fact that one is able to come up with reason or reasons for

[109] We should expect Abel to react the same way if the distributive principle that the lottery scheme picks out is his least favored principle, i.e. the PEC.

compelling Mabel to accept the principle. I am only saying that such reason or reasons might not be justifiable in the light of Mabel's expected utility.

Perhaps the best Abel can do to get Mabel to go along with him is to propose some rotational policy; a policy that alternates among the three principles of distribution. Under this policy, both of them would agree to support a scheme of cooperation that distributes the cooperative surplus by rotating, say, yearly, the three principles of distribution. This seems a viable and reasonable way of securing Mabel's, as well as Abel's continued acceptance of the terms of the contract. In settling for a rotational policy however, all principles of distribution are made equally valid and legitimate. But if all principles of distribution are equally legitimate, then Gauthier seems mistaken to claim that the MRC principle is a unique bargaining principle.

There is a sense in which I have ignored Gauthier's claim that we ought to recognize the contribution or role played by 'agreement' in the production of the cooperative surplus. Gauthier has consistently argued that the attractiveness of the MRC principle over alternative principles of distribution is that it recognizes the role played by agreement in securing the cooperative surplus. Agreement, he says, is both necessary and sufficient in cooperative activities and this justifies the choice of the MRC rule over other principles. Gauthier says this about the role of agreement in the money market investment of Abel and Mabel:

> As long as his [Abel's] agreement is both necessary and sufficient to double the return on any dollar invested by Mabel, then he can bargain with her over the distribution of that increase in return, and reasonably claim half. The size of their two contributions, either in dollars or in

> relation to each other, as long as they are within the limits laid down by the requirement that Abel's agreement be necessary and sufficient, is immaterial.[110]

The kernel of Gauthier's argument is that without the agreement of Abel, there would be no cooperative surplus. If Abel had refused to cooperate with Mabel, Mabel would have been limited to the $30 she makes on her own; hence, she must recognize the role played by Abel's agreement in generating the cooperative surplus, i.e. the additional $50. Therefore, it is reasonable to believe that MRC will command Mabel's consent because in contrast to PP and PEC, it recognizes and factors the contribution of agreement into the distribution of the cooperative surplus.

But Gauthier's argument for the role played by agreement in producing the cooperative surplus is misleading. Why must we focus exclusively on the agreement of Abel and not that of Mabel? Why shouldn't we give equal weight to the role played by both of their agreement in generating the cooperative surplus? In particular, since the production of the cooperative surplus requires agreement from both of them why should Mabel be the one to concede more and receive less? Under the MRC rule, Mabel concedes $55 and receives $25 of the cooperative surplus, while Abel concedes $45 and receives $25 of the cooperative surplus. If we assume that MRC, PP and PEC are the only principles of distribution, and given that Abel's agreement as well as Mabel's is necessary to the production of the cooperative surplus, then we have no reason to assume that they would accept as fair any principle that discounts either person's agreement.

---

[110] Gauthier, "Moral Artifice: A Reply by Gauthier," p.391.

If we take seriously the role played by Abel's agreement as well as Mabel's, then it would seem that they would choose a principle of distribution that takes concession at mid-point. In which case, that will be the PP principle, under which Mabel concedes $60, i.e. she receives $30, and Abel concedes $40, i.e. he receives $20. Alternatively, since agreement from both is necessary to generating the cooperative surplus they might consider a new principle of distribution, a principle that averages all three principles of distribution or concession points. In this case, Mabel would receive 29.45 (the average of 33.34 + 30 + 25) and Abel will get 20.55 (the average of 25 + 20 + 16.66).

In any case, because Gauthier's argument places more emphasis on the role agreement plays in generating the cooperative surplus it seems to discount the role played by actual contributions. Suppose that the contract that Abel and Mabel are to bargain for is not money market investment but sentinel duties. Suppose also that their contributions are not money but hours. Suppose as well that predatory activities in the neighborhood in which they live require 24-hour surveillance for seven days a week. And suppose finally that Abel cannot engage in any sentinel duties exceeding 10 hours a day, a situation that requires Mabel to cover a 14-hour sentinel duty. Unquestionably, individually, Abel and Mabel cannot secure the neighborhood alone so they must necessarily cooperate. In this example, as with the money market investment example, Mabel contributes more than Abel. If we accept Gauthier's argument on the role played by Abel's agreement, then Mabel ought to accept the arrangement even though she contributes 4 more hours than Abel since the size of their "contributions, either in dollars or in relation to each

other, as long as they are within the limits laid down by the requirement that Abel's agreement be necessary and sufficient, is immaterial."[111]

To see the deceptive force of this argument, let us suppose further that Mabel's 14-hour sentinel duties prevent her from taking an extra work shift; each work shift is 4 hours and pays $15 an hour. If we assume that (1) both of them work a 4-hour shift each day and (2) Mabel and Abel respectively spend 6 hours and 10 hours for sleep and other house, social and recreational activities every day, then each makes $1,825 monthly ($21,900 annually). There is surely a sense in which this arrangement is unfair to Mabel because her 14-hour sentinel duties prevent her from making an extra $1,825 monthly (for a 4-hour work shift). If we follow Gauthier's argument about agreement and the production of the cooperative surplus, then Mabel must surely accept this arrangement. But this is definitely unfair to Mabel and she would be unreasonable to accept the MRC principle, and accordingly Abel's justification. Is there a way of defending Gauthier's view regarding the uniqueness of MRC? Can we show that MRC is an appropriate and legitimate principle of distribution?

3.1.1.2 MRC, Inequalities, and the Archimedean Standpoint

The point of view of the Archimedean chooser might provide us with a perspective from which to defend the MRC principle of distribution not just as an appropriate distributive principle but also as a unique one. The point is that the stance of the Archimedean chooser might give Mabel a valid reason to pay attention to Abel's

[111] Ibid, p.391.

justification of the MRC principle. If she considers that the principles they both choose to govern their behavior are constrained by the 'Archimedean stance'[112] then she may have a legitimate reason to support a scheme of cooperation circumscribed by the MRC principle of distribution.

Why should we expect that both would favor the MRC principle or the conditions or scheme of cooperation that it makes possible if we assume that they choose bargaining principles from the Archimedean standpoint? The reason is that from the Archimedean standpoint rational actors are mostly interested in selecting principles that discount unjustifiable inequalities, i.e. they are primarily moved to choose principles that promote an essentially just society. We might assume that Abel's situation, particularly his inability to engage in any sentinel duties exceeding 10 hours, is a kind of inequality—perhaps he has some health issue for which sleeping more hours than Mabel is justified. Examples of unjustifiable inequalities that Gauthier discusses are the right of bequest (which I shall discuss in section 2.2.3), factor rent, and the socialization process whereby males are "encouraged to actualize capacities repressed in females."[113]

Unjustifiable inequalities are to be discounted because they introduce unfairness into the contract, and, as we know, removing unfairness from the contract, for Gauthier, is one of the aims of a rational morality. Impartiality is satisfied just in case rational agents in choosing as Archimedean choosers choose in full view of everyone's individuality (human capacities, preferences, and circumstances) that is, they choose not as if they were *this particular person*, but as

[112] Gauthier defines the Archimedean point in moral theory as "that position one must occupy if one's own decisions are to possess the moral force needed to govern the moral realm." MbA, p.233.
[113] Ibid, p.263.

if they *were every person*. This is different from Rawls' idea of impartiality. Rawls, as we have seen, takes impartiality to be satisfied when rational agents in adopting the rule of maximin, choose principles behind the veil of ignorance and from the standpoint of the least advantaged member of society.

To mitigate unjustifiable inequalities, Gauthier claims that agents would prefer, in general, that their expected share of the cooperative surplus be related not to what they *actually* contribute since their actual contribution may reflect the contingent permissions and prohibitions found in any social structure, but to what they *would* have contributed in a feasible social arrangement or scheme of cooperation. Gauthier says this about the relationship between contributions and the capacities and character traits of agents:

> Each person's expected share of the fruits of social interaction be related…to the contribution he would make in that social structure most favorable to the actualization of his capacities and character traits, and to the fulfillment of his preferences, provided that this structure is a feasible alternative meeting the other requirements of the Archimedean choice.[114]

By relating either of their benefits to their relative concession point, the MRC principle recognizes the importance of remedying unjustifiable inequalities and hence, without the use of force, MRC would more than likely command the assent of both Mabel and Abel.

Gauthier's reasoning seems to be motivated by his belief that fairness and justice in the hypothetical contract are determined largely by the extent they remedy unjustifiable inequalities. To remedy unjustifiable inequalities and ensure

[114] Ibid, p.264.

mutual benefits, Gauthier believes returns should reflect contributions and agreement. Returns reflect contributions and agreement as long as they reflect social contingencies of socialization and inequalities. This is not because this particular social structure fails to relate benefits to contributions or allows individuals to take advantage of their fellows but because "it fails to relate benefits to the contributions each person would have made had each enjoyed similar opportunities and received similar encouragements."[115]

It is an incontrovertible fact that NFL and CFL players handsomely benefit from playing football; they make lots of money and are famous in all sorts of ways. Equally true is that they eke out their income and get all of the attendant fame by plying a profession that encourages getting 'in your face' and 'aggressive' most, if not all, the times. The hard hits, the wild turns and runs, the sacks, the strong tackles, the stunts, the aggressive punts, the interception and touchdown catch or runs are all part of this 'in your face'—all of which we may suppose are signs of 'appropriate machismo.' We might loosely compare playing a professional football game without any of the 'aggressiveness,' to playing chess without the King, in the sense that each seems 'ungainly.'

The 'aggressiveness' or machismo displayed by football players is sometimes the sort of behavior that is repressed in females. Females are often socialized into behavioral patterns that make it difficult, if not impossible, for them to be the sort of people fit for the NFL and CFL. Yet, we require them to *bargain* with males, who are "encouraged to actualize capacities repressed in females," who in exploiting this situation get to benefit more than females. Now, if part of the

[115] Ibid, p.263.

reason why females do not play in the NFL, or if part of the reason why there is no professional female football league, or if part of the reason why female football players do not attract as much money as their male counterpart is that they are not able to exhibit appropriate machismo, then given that our socialization process is the reason for their behavior, they ought to be compensated. This is exactly what the MRC principle does.

Suppose in our investment example it is Mabel and not Abel that has to justify the MRC principle of distribution because it is the principle she favors. Abel is a football player and Mabel is not, perhaps she is a teacher who is not well paid. Abel favors PEC and contributes more than Mabel. Since Mabel's contribution would have been more were she to play in a football league that rewards her as much as Abel or close to what Abel earns, it would be right to conclude that any scheme of cooperation engendered by any principle other than the MRC rule would be unfair to her. This would seem right if we consider that Mabel, as well as Abel, bargain and choose principles from the Archimedean point, the standpoint of impartiality. The impartial Archimedean standpoint is satisfied if the cooperative surplus is divvied up based on what Mabel (or each) would have contributed "had each enjoyed similar opportunities and received similar encouragements" to actualize and develop any 'machismolike' or appropriate capacities.

That Gauthier moves towards this direction is evident from his consideration of factor rent—"the premium certain factor services command, over and above the full cost of supply, because there is no alternative to meet the

demand."[116] In arguing against factor rent as undeserved, as in the case of Wayne Gretzky (Gauthier's version of Nozick's Wilt Chamberlain), Gauthier directs us to the role social interaction play in making factor rent possible. He writes, "The benefit represented by factor rent is part of the surplus afforded by that enterprise, for it arises only in social interaction."[117]

It is a fact that females do not play in the NFL and CFL. It is equally true that they do not have a professional football league that rewards them in the same manner as their male counterpart. We may suppose that this is due partly to our socialization process. By this, I mean the social process of encouraging males and females to develop and hone different sorts of traits and physical abilities. Since females do not have a professional football league that rewards them in the same manner as their male counterpart, we do well to rectify the inequalities that arise by imposing a "confiscatory tax on rent"[118] by collecting that portion that male football players charge over and above the full cost of supplying their skills and abilities.

Jean Hampton has criticized Gauthier's argument against factor rent. She argues that Gauthier conflates hockey-related opportunity costs and non-hockey alternatives for Gretzky and this conflation leads to his implausible rejection of factor rent.[119] Noteworthy here is that Gauthier argues against factor rent because he considers it a paradigmatic example of an unjustifiable inequality and Hampton's contention is that Gauthier is mistaken to consider factor rent as an

[116] Ibid, p.272.
[117] Ibid, p.274.
[118] Ibid, p.273.
[119] See Hampton, "Equalizing Concessions in the Pursuit of Justice," pp.156-157.

unjustifiable inequality because there is a significant difference between hockey-related opportunity cost and non-hockey alternatives. My concern is different from Hampton's. I am interested in examining how far-reaching Gauthier's argument against unjustifiable inequalities goes. Specifically, I am interested in his overarching view of rectifying unjustifiable inequalities.

To the extent that Mb(CM)A focuses on possible opportunities and contributions, Gauthier can indeed make a case for "social minorities" and "marginalized groups" like women. In fact, the assurance of circumscribing interactions by the constraints of justice, Gauthier argues, "meets the concern emphasized in feminists thought, that sociability not be a basis for exploitation."[120] Having said this, one wonders if Gauthier's interest in rectifying unjustifiable inequalities goes far enough given that in his analysis of factor rent and the right of bequest he limits inequalities that arise from them to the outcome of socialization. Specifically, if Gauthier considers factor rent and the right of bequest as wedded to socialization processes and hence to unjustifiable inequalities that need to be remedied by the contract, one wonders if he is justified in excluding the inequalities that arise from natural endowments, namely, the sort that Rawls thinks are fundamental in determining the trajectory that the lives of people take.

Given what the NFL or CFL offers, it seems to matter much for an individual to play in these leagues. For indeed, many males would reasonably want to, but if one lacks the stature, physical abilities, and skills required in the NFL and CFL, "maleness" would not make a significant moral difference. Some of these abilities and skills, which are honed and actualized in society are "attached" to

[120] MbA, p.351.

people as part of their natural endowment. If Gauthier believes we should recognize unjustifiable inequalities such as factor rent and others occasioned because certain groups of people are "mis-socialized" or misshaped, why should we not consider factors such as those of nature that place some people on the wrong side of things by "mis-bestowing," or "withholding" certain natural endowments from them, more especially when these significantly affect what careers they decide to pursue. It would seem that from a moral point of view, both sets of inequalities are morally arbitrary and should be discounted by an appropriate theory of justice.

We can imagine a person trying to pursue a career arrives at the disquieting conclusion that given his or her stature and ability, he or she can only be a garbage collector or janitor. This individual's reasoning that he or she is not a candidate for the NFL and CFL or indeed the NBA might proceed in the following manner:

> **P1**: I desire to play in the NFL, CFL or NBA
> **P2**: I cannot play professional football because I lack the stature, physical abilities and skills required in the NFL and CFL.
> **P3**: I also cannot play in the NBA because I am too short and fragile.
> **Conclusion**: Therefore, I will become a garbage collector or janitor, which does not require anything considerable from me that I cannot satisfy and follow through on.

Gauthier's view on rectifying unjustifiable inequalities, it would seem, commits him to support the garbage collector's or janitor's push for some form of redress or rectification. To be consistent Gauthier has to argue that all forms of inequalities—whether social or natural—ought to be addressed and compenstated. After all,

when rational actors choose as every person from the Archimedean standpoint, they are interested in selecting principles that discounts unjustifiable inequalities.

Since inequalities of any sort, *are inequalities all the same*, there is no reason to assume that rational actors in the Archimedean standpoint, who are at least "aware that [they have] an identity"[121] would pick which inequality society ought to rectify. Rationally, it would seem that they would opt for a social scheme where all inequalities are rectified since they would be bargaining not as *this* person but as *each* person. If this is the case, then, one can logically question whether or not Gauthier's recognition of these inequalities and the method of remedying them is significantly different from Rawls' much criticized "lexical difference principle."

Yet, Gauthier can insist that given the difference between social and natural inequalities, he is justified in decoupling them and focusing on the former. He could object by querying our analysis of the social and natural dimensions of inequalities. Particularly, he could question the rationale behind our conflating the natural with the social. Inequalities caused by the socialization process are *social*, and inequalities caused by lack in certain natural endowments—such as that of the garbage collector or janitor—are *natural*. Different things cause both set of inequalities, hence we should treat them differently. We might hold society morally responsible for social inequalities and require that it rectify the inequalities it brought about, but we cannot hold nature or indeed society morally responsible for the inequalities that nature brings about and require that it (or society) rectify them.

[121] Ibid, p.251.

There is surely a difference between "natural forces" such as a Tsunami or an earthquake "conspiring" to destroy my house and social policies "conniving" to pull down my house. In the same way, there is a difference between social policies preventing one from living in a certain neighborhood and gravity or the certain lack thereof in natural resources preventing one from living on Mars. Whereas social events and inequalities seem preventable, natural events or inequalities do not seem preventable. Given the fundamental difference between the social and the natural, and hence between social and natural inequalities we would expect any principle of distribution to recognize the difference. A distributive principle recognizes the difference by rectifying those that are consciously brought about through the actions of humans and society.

3.1.1.3 The Archimedean Perspective versus the Individual Perspective

Before I move to the second sub-theory there is one issue concerning the theory of rational bargaining that I want to discuss. It is the criticism that the ideal choice from the Archimedean point weakens Gauthier's attempt to derive moral constraints from individual rationality. I turn to Hampton for a statement of this objection.

> But more worrying is the fact that [Gauthier's] shift in methodology undercuts his Hobbesian approach to generating moral constraints from individual rationality. The social contract which assumed determinate people who bargained with knowledge of their identity and factor endowments was supposed to demonstrate to us determinate individuals why the adoption of 'moral principles such as the MRC rule are also

> individually advantageous for us. But a determinate individual is probably not going to find individually rational the adoption of constraints agreed to by 'proto-people'.[122]

Hampton's point is that earlier in chapters V, VI, and VII of MbA, Gauthier specified the contract situation in terms of determinate, fully socialized individuals who have utility functions and factor endowments and who know what their utility functions would be under various schemes of cooperation. But in chapter VIII, the Archimedean chooser "simulates a bargain among people who select a scheme of cooperation not on the basis of who they *are*, but on the basis of *who they could be* in any of these schemes."[123] This modification of Gauthier's contract method indicates a shift in his conception of the purpose of the contract methodology. Previously, Gauthier claims that the aim of the methodology is to select principles for individuals to use in order to promote and ensure a desirable cooperative relationship; the contract is to govern individual morality, which sets apart his contract methodology from the contract methodology of Rawls, whose approach is meant to govern basic social institutions. But in chapter VIII, Gauthier, like Rawls, employs the contract method to choose principles that are for the structuring of social institutions and systems that play a profound role in creating individuals.

Hampton's objections are less convincing than they appear when we consider what Gauthier says about the aims of the methodologies in chapter VIII. He says this towards the end of the chapter, "Moral theory offers an Archimedean point analysis of human interaction. The theory of rational choice offers an analysis

[122] Hampton, "Can We Agree on Morals?" in *Canadian Journal of Philosophy*, vol. 18, 1988, p.352.
[123] Ibid, p.351, emphasis in original.

from the standpoint of each interacting individual."[124] Whereas the individual with reference to rational choice theory is concerned with rational agreements as they affect "individual interactions," the Archimedean chooser with reference to the ideal choice is concerned with impartial choices as they affect "human interactions." The bottom line is that for Gauthier, the impartial perspective of the ideal rational actor must cohere with the perspectives of rational individuals engaged in strategic choices, and their coherence is a demonstration of the appropriateness of both methodologies.

The individual rational chooser knows who she is and what her preferences, utility functions and circumstances are. She chooses a particular social structure or scheme of cooperation in full view of her capacities, preferences, utility functions and circumstances. The Archimedean chooser does not know what her identity is, i.e. she does not know if she is this or that person. She however knows she has an identity. She also knows the capacities, preferences and the utility functions, and circumstances of all the rational choosers, one of whom she will turn out to be. The ideal chooser chooses on behalf of every chooser a scheme of cooperation and gives no special favor to any existing arrangement or to any chooser. She proceeds systematically by taking up the perspective of each rational chooser one by one and considering what bargain fixing on a scheme of cooperation that chooser would willingly agree to as a utility-maximizing chooser.

The Archimedean perspective does not itself yield compliance with the social contract since it primarily serves to confirm "from a moral perspective, the

[124] MbA, p.266.

rational derivation of impartial constraints on straightforward maximization."[125] Both methodologies in this sense yield the same results, as far as principles, social institutions and practices are concerned. For example, an individual who chooses rationally "may consider existing social institutions and practices ultimately unjustified." But that same individual may agree that given the person she really is, the existing social institutions and practices "afford her a fair share of the benefits of social cooperation. For the person she is may not be the person that, she supposes, she would have been, in an essentially just society."[126]

3.1.2 Constrained Maximization and the Problem of Rational Compliance

Constrained maximization, Gauthier says, is one of the most important, if not, the most important components of Mb(CM)A. It is an optimality-enabling strategy or policy that identifies rationality with utility maximization at the level of dispositions to choose and which links "the idea of morals by agreement to actual moral practice."[127] Its importance is evident in the role it plays within the general problem of rational compliance, which is essentially the PD.[128] Constrained maximization (CM) is a strategy that requires us to be disposed to mutually advantageous moral constraints or to choose to cooperate provided others are so disposed or choose to cooperate as well. CM does not identify with directly utility-maximizing actions; rather, it identifies with the dispositions to cooperate. The

[125] Gauthier, "Moral Artifice: A Reply by Gauthier." p.414.
[126] Ibid, p.415.
[127] MbA, p.168.
[128] For Gauthier's discussion of the significance of CM to morality see Gauthier, "Uniting Separate Persons," in *Rationality, Justice and the Social Contract: Themes from **Morals by Agreement***, Gauthier and Robert Sugden (eds.), Ann Arbor, University of Michigan Press, 1993, pp.185-191.

conditional aspect of CM effectively distinguishes it from straightforward maximization, i.e. the other strategy that Gauthier discusses.

Straightforward maximization requires that we choose the action which yields us the greatest expected utility given our expectations about the actions of those we are interacting with. Unlike CM, which identifies with the dispositions to cooperation, straightforward maximization (SM) identifies with directly utility-maximizing actions and requires that we choose those actions that maximize expected utility. SM, Gauthier claims, is not a rational strategy to adopt because it leaves us with limited opportunities for cooperation, suboptimal outcomes, and fewer utilities. CM, on the other hand, is a rational strategy to adopt because it provides us with more opportunities for cooperation, optimal outcomes, and more utilities. On average, optimal outcomes are of greater utility to an individual than equilibrium outcomes. This is because in any given optimal outcome an individual will never be any worse off than she would be at any equilibrium outcome.

Notice that Gauthier parses the difference between CM and SM strategies in terms of individual and joint strategies. Those who have a disposition to CM, i.e. a constrained maximizer (we will simply call these CMers) aim for mutually advantageous outcomes and they have the disposition to act on joint strategies. An individual, Gauthier says, is a CMer if that individual "seeks in some situations to maximize her utility, given not the strategies but the utilities of those with whom she interacts."[129] In contrast, an individual is a straightforward maximizer (SMers, for short) if that individual seeks "to maximize [her] utility given the strategies of

[129] MbA, p.167; A CMer, Gauthier says, "base[s] actions on a joint strategy, without considering whether some individual strategy would yield her greater expected utility," p.167.

those with whom [she] interacts.[130] Since SMers have the tendency to act on individual strategies, cooperation for them might tend to be exploitative rather than beneficial. Because the Foole or rational skeptic is prepared to violate cooperative arrangements whenever it yields him the greatest expected utility he or she accepts the rationality of SM.

A CMer is disposed to cooperation insofar as others are similarly disposed. Cooperation results when we employ a CM or joint strategy. If it is the case that to agree to cooperate is to agree to moral constraints, then there is cooperation only when we employ a CM strategy. It is pointless for you and me to agree on a principle of distribution or on how to run a scheme of cooperation if we are not prepared to honor the terms of the agreement setting up such a scheme. We constrain our behavior when we honor the terms of an agreement and to honor the terms of an agreement is to cooperate. Given then that we maximize expected utility when we cooperate, our employing a CM-joint strategy would seem rational. Note, however, that the disposition to CM or the disposition to cooperate is a conditional one. It is conditional on (a) the expectation that others will be disposed to cooperate as well, and (b) on the expectation that cooperative arrangements will yield greater utilities than noncooperative arrangements.

There are two kinds of CMers, according to Gauthier. There is broad compliant and narrow compliant. An individual is a broad compliant if that individual is "disposed to cooperate in ways that, followed by all, merely yield her some benefits in relation to universal non-cooperation." And a CMer is a narrow compliant if the CMer is "disposed to cooperate in ways that, followed by all, yield

[130] Ibid, p.167.

nearly optimal and fair outcomes."[131] The difference between these dispositions is that they have different thresholds for cooperation. A broad compliant and not a narrow compliant is willing to lower her utilities to facilitate cooperation.

A narrow compliant, Gauthier says, is the model of practical rationality and generally does better than a broad compliant in cooperative arrangements because she is "prepared to be co-operative whenever co-operation can be mutually beneficial on terms equally rational and fair to all."[132] Her dispositions prevent others from exploiting her. Conversely, the disposition of a broad compliant to act on cooperative arrangements yielding minimal benefits exposes her to exploitative activities, as others easily "maximize their utilities at her expense by offering 'co-operation' on terms that offer her but little more than she could expect from non co-operation." [133] A broad compliant is closer to a de facto altruist. Others exploit and maximize their benefits at the expense of a broad compliant. Similarly, others take advantage of and improve their position at the expense of a de facto altruist.

According to Gauthier, in order to be able to form the disposition to cooperation, it is important that agents have complete knowledge not only of their preferences, life-plans and characteristics but also of all feasible social arrangements as well as the one they happen to favor. To deprive them of full knowledge of the various social arrangements and the different capacities that they possess, as Rawls does via the veil of ignorance in JaF, is to encumber the bargaining process. This is because for the bargaining process to achieve its purpose of producing acceptable outcomes—and for it to enable a bargainer to

[131] Ibid, p.178
[132] Ibid, pp.178, 179.
[133] Ibid, p.178.

make the right concession to others—a bargainer must be able to evaluate all possible life plans and feasible social arrangements, including the initial factor endowments she and others possess and the role they place in producing the cooperative surplus. If we take the social contact as one of rational bargaining, then the initial position of bargainers must be determinate since the bargain can only proceed from a condition of complete knowledge of the initial factor endowment each person brings to the bargaining table.

When bargainers choose principles as individuals, they choose principles that reflect the particular circumstances of everyone. But they can only choose principles that reflect the particular circumstances of everyone if they have knowledge of the particular abilities and circumstances each one has. And when they choose principles from the Archimedean standpoint, they choose principles that reflect the particularity of each person, since they choose as each person. When bargainers choose principles as Archimedean choosers the standpoint prohibits them from skewing the principles to *any particular* individual and circumstances. In either case, whether bargainers choose principles as individuals or they choose as Archimedean actors, the individuality of each person is, has to, and must be accessible to every chooser. Although Gauthier introduces the Archimedean perspective into the social contract as an addition to the individual perspective of the rational chooser, it is the latter perspective that is basic to the contract because it is the perspective that yields compliance with the contract. The Archimedean

perspective primarily serves to confirm "from a moral perspective, the rational derivation of impartial constraints on straightforward maximization."[134]

It is important to emphasize that both methods of selecting principles is different from the method of selecting principles in JaF (See 3.1.1.3 The Archimedean Perspective versus the Individual Perspective). If we consider the individual perspective, the individual that chooses in Mb(CM)A identifies with his or her individual characteristics and circumstances, but the individual that chooses in the JaF does not identify with his or her individual characteristics and circumstances, but rather identifies with the least well-off group. And if we consider the Archimedean perspective, the individual that chooses as the Archimedean chooser in Mb(CM)A "is not aware of her identity, [however] she is aware that she has an identity."[135] She identifies with the individuality of every person, but the individual that chooses in the JaF, even though she chooses impartially, identifies with the least well-off group.

There is a sense however in which both Mb(CM)A and JaF are similar. Both theories are interested in anchoring the principles of justice on a sound footing—on impartiality and justice. Rawls goes about achieving this by employing the veil of ignorance to screen off any identifying mark of each person. The real individual, who must identity with the ideal choice in JaF, is ignorant of her true identity (capacities, talents, attitudes, preferences) as well as the true identities of others. Ignorance of one's identity guarantees that one takes special

[134] Gauthier, "Moral Artifice: A Reply by Gauthier," p.414.
[135] Ibid, p.251.

care to prevent the worst befalling one and, thus, in taking on this perspective one ends up showing concern for all.

Gauthier, on the other hand, goes about achieving fairness in the contract by placing decision-makers outside the veil of ignorance, either as individual rational choosers or as Archimedean choosers. For the rational actor in choosing as an individual rational chooser chooses in complete view of her individuality, and in choosing from the Archimedean standpoint she chooses not as if she has an equal chance of being each of the persons affected by her choice but as if she *were* each of those persons. In maintaining the separate identities and utilities of persons, she chooses as if she were bargaining as each person. The choice of an ideal actor is the choice of *every person*.

In addition to possessing complete knowledge of various social arrangements and the different capacities that they possess, decision-makers also possess information about the dispositions of other rational actors. The rationality of CM is not completely insensitive to the identity of decision-makers. It is rational for us to follow a strategy of CM only when we are interacting with those whom we believe to be other CMers. This is important since it enables decision-makers to avoid interactions that are exploitative as well as those that are disposed to exploit for personal benefit. To avoid exploitative interactions CMers are required to place themselves in a position that will enhance their ability to predict sufficiently well the strategy and dispositions of other decision-makers.

Gauthier places his agents in this position. He argues that, for the most part, agents are able to predict sufficiently the dispositions and strategy of others

because the dynamics of the bargaining process and the contract promotes and encourage *translucency* rather than opacity. His argument is that rational actors might find some advantage not only in not masking their true intentions, but also in finding some ways to guarantee and convince others that they are not masking them. In not masking their dispositions, they save themselves the trouble of being 'misread' as cheats and as SMers, and from having to settle for noncooperative, equilibrium or less than equilibrium outcomes instead of reaping the benefits of cooperative, optimal outcomes. I say more about this in a moment

Thus far, I have presented an outline of how Mb(CM)A negotiates the rationality of compliance, which I indicated, is grounded on the distinction between SM and CM strategies. SM, Gauthier argues, leaves SMers worse off because it is a strategy that identifies only with directly utility-maximizing actions. By contrast, CM leaves CMers better off because it is a strategy that identifies with the disposition to cooperate. Central to the claim that a CM strategy leads to optimal outcomes is the thought that because, on average, we are sufficiently able to identify both dispositions and what dispositions decision-makers happen to possess, there is a tendency for CMers to do better than SMers because they will "obtain cooperative benefits that are unavailable to straightforward maximizers."[136] On this view, an SM strategy is a costly strategy. As Gauthier puts it, whereas those disposed to CM 'will be welcome partners in mutually advantageous cooperation, in which each relies on the voluntary adherence of the others, those disposed to straightforward maximization will be excluded.'[137]

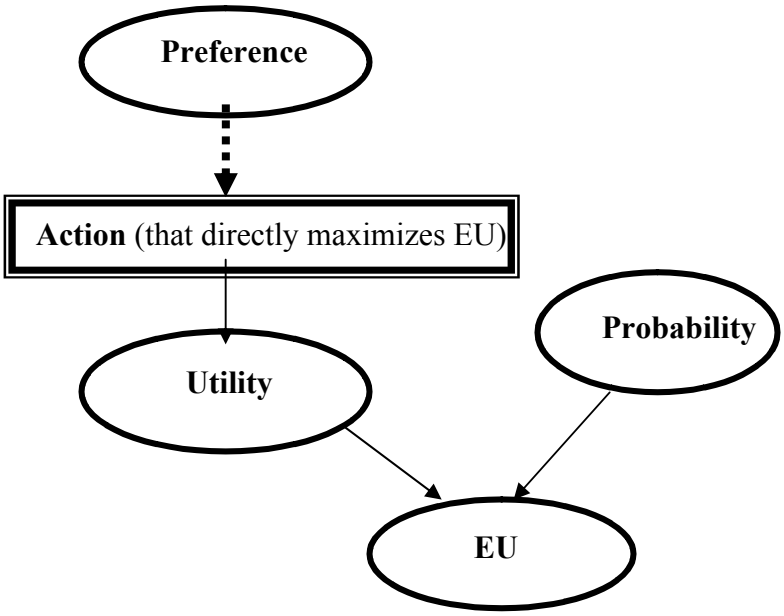
[136] Ibid, p.170.
[137] Gauthier, "Why Contractarianism?" in *Contractarianism and Rational Choice*, p.25.

The rational skeptic is amoral because he does not constrain his behavior. He acts on an SM strategy by identifying only with directly utility-maximizing actions. In pursuing his concerns and interests whatever they may be, without constraints, he undermines the foundation of morality for which constraints are essential. In contrast, a CM strategy opens the space for morality because it encourages us to constrain our behavior, even in those situations where we may not seem to be maximizing expected utility. A CMer and not an SMer thus proves to be the rational *qua* moral individual.

3.1.2.1 CM, Rationality, and the Theory of Rational Choice

To understand Gauthier's argument for the rationality of CM we need to understand what he thinks is misleading and mistaken about the notion of rationality that the received view of rational choice theory advocates. The received view of rational choice theory (call it TRC) identifies with the rationality of SM. According to TRC, we choose the best action or strategy according to our stable preferences and the constraints we face. Rationality, for TRC, requires us (a) to choose directly utility-maximizing actions when those actions maximize EU, and (b) to maximize utility, given the strategies of those with whom we interact. On this view, an action is rational if it is EU-rational. An action is EU-rational if and only if that action maximizes EU or if it offers an agent an expected utility not less than any alternative action. Figure 3.2.2.1a illustrates this.

Figure 3.1.2.1a: **Actions and Rational Choice Theory**

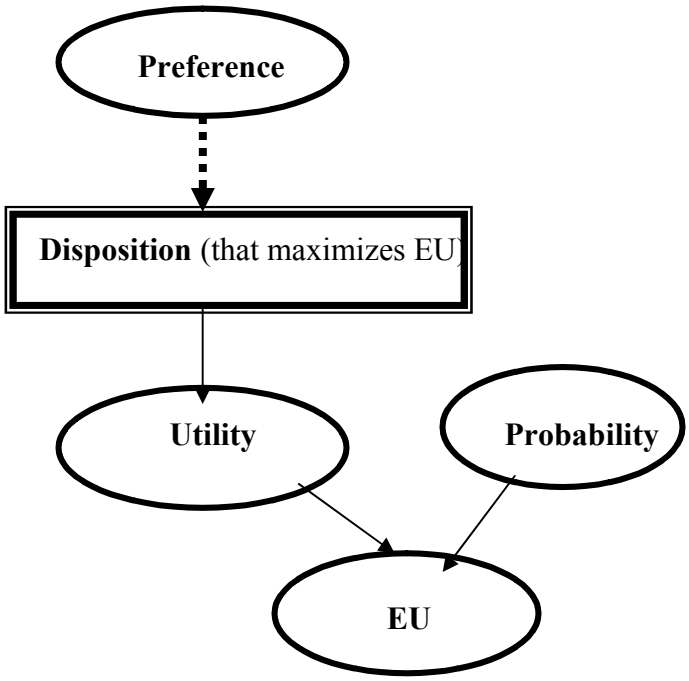


An action refers to the options available to an agent, while utility is the measure of an agent's relative satisfaction from consuming a variety of baskets of goods. Utility captures individual preferences—desires and beliefs—in the sense that it is a measure of our coherent considered preferences about outcomes, and as such, it remains subjective, hence it is sensitive to context. If "utility is a measure of an agent's preferences over possible outcomes, and so derivatively over that agent's possible actions,"[138] then as figure 3.1.2.1a shows, our reasons for acting are directly linked to whatever preferences we happen to have and to those actions that directly maximize our utility functions. An action is therefore rational, on TRC's view, if choosing that action gives us greater expected utility than would choosing any alternative action.

[138] Gauthier, "Uniting Separate Persons" in *Rationality, Justice and the Social Contract: Themes from* ***Morals by Agreement***, p185.

If we accept the rationality of CM then we must reject TRC's view of practical rationality. Because CM and not SM, according to Gauthier, allows us to maximize expected utility he proposes a modification of TRC that he believes is consistent with a CM strategy and congenial with an agent's utility function. Gauthier's modified view (let us call it MTRC) is consistent with the rationality of CM in the sense that it introduces dispositions into rationality, where the dispositions speak to the rationality of constraints. MRTC is the view that we choose those dispositions that help us to maximize EU according to our stable preferences and the constraints we face. Rationality, on this view, requires us to act on (a^1) dispositions, where the dispositions maximize expected utility and (b^1) maximize utility in some situation, given not the strategies but the utilities of those with whom we interact. This is illustrated in figure 3.2.2.1b.

Figure 3.1.2.1b: **Dispositions and Constrained Maximization**



For MTRC, therefore, a disposition is rational if and only if an agent expects to do better holding such a disposition than any alternative disposition. Gauthier puts it this way:

> We identify rationality with utility-maximization at the level of dispositions to choose. A disposition is rational if and only if an actor holding it can expect his choices to yield no less utility than the choices he would make were he to hold any alternative dispositions."[139]

Gauthier provides the following example to illustrate the importance of dispositions, particularly CM dispositions or dispositions to cooperation:

> Suppose for example that I promise to assist you next week in some way provided you assist me now. I reasonably expect to benefit more from so promising than from any alternative open to me, since I have no alternative way of gaining your assistance, and its benefit to me is much greater than the cost of reciprocating. Suppose you accept my promise and in consequence assist me. Next week I may have no directly utility-based reason to assist you; neither reputation effects nor prospects of future interaction between us need outweigh the costs of actually giving you the promised assistance. Nevertheless, I may expect to be doing better honoring my promise than I should be doing had I made no promise – and in consequences, not received your assistance. And given this expectation, my promise affords me sufficient reasons to carry it out – not of course according to the received theory of rational choice, but according to the alternative theory that embraces constrained maximization.[140]

[139] MbA, pp. 182, 183.
[140] Gauthier, "Uniting Separate Persons," p.186.

In this example, as it is with the babysitting example I discussed in chapter one, the expectation to do better by promising to reciprocate provides sufficient reason to constrain behavior and to honor commitments. For an SMer, the focus is on directly utility-maximizing actions. He has no directly utility-based reason to honor commitments. Although you babysat for me last week, as an SMer, I am not going to babysit for you this week because I have no directly utility-based reason to do so. The SMer, like the rational skeptic *qua* freerider chooses defection in the PD. However, for a CMer, the focus is on those dispositions that provide us higher utilities. Such a person honors commitments because of the expectation of doing better by honoring commitments. As a CMer, the reason you babysat for me is that you expect to do better by keeping your commitment than you would have done if you have not made the commitment to babysit for me in the first place.

Gauthier's modification of TRC's notion of rationality[141] shifts the focus of reasons for acting away from directly utility-maximizing actions to dispositions. Under TRC, because the individual chooses directly utility-maximizing actions, she acts on reasons directly related to her utilities, but under MTRC, because the individual chooses dispositions, she acts on reasons indirectly related to her utilities. The individual, who acts on reasons indirectly related to her utilities, does better than the individual who acts on reasons directly related to her utilities. The latter is an SMer and the former is a CMer. Reasons that indirectly relate to one's utilities include the execution of a plan or the honoring of a commitment. In executing a plan or honoring a commitment the person knows that she would not

[141] When I use 'the theory of rational choice' or 'rational choice theory,' note that I am referring both to TRC and MTRC. And when I use TRC, I am specifically referring to the received view of rational choice theory.

be maximizing EU at that time. But since the plan or commitment had been rationally undertaken (because it maximizes EU), and given that in executing or honoring the plan she would expect to be doing better (in those terms) than she would be doing had she not undertaken it she would honor it.[142]

Two important differences between TRC and MRTC[143] are obvious from the foregoing. First, TRC is exclusively concerned with actions, while Mb(CM)A is exclusively concerned with dispositions, and is interested in actions to the extent dispositions relate to them. Second, TRC takes the reasons for acting exclusively and directly from one's utilities, whereas Mb(CM)A takes the reasons for acting from reasons that are indirectly related to one's utilities. These differences notwithstanding, Gauthier's modification of rationality is still EU-focused. On Mb(CM)A, as it is on TRC, the reasons for acting are strictly provided by or dependent upon expected utility.

By associating the reasons for acting or circumscribing rationality by expected utility, the theory of rational choice excludes from the purview of rationality all 'moral reasons' not found in an individual's utility function. Particularly, that theory leaves out from rationality reasons that appeal to value, i.e. reasons that are from an agent's point of view significant in determining what actions that agent chooses. And as we would see in later chapters of this project, these reasons are not only important in understanding how an agent chooses in

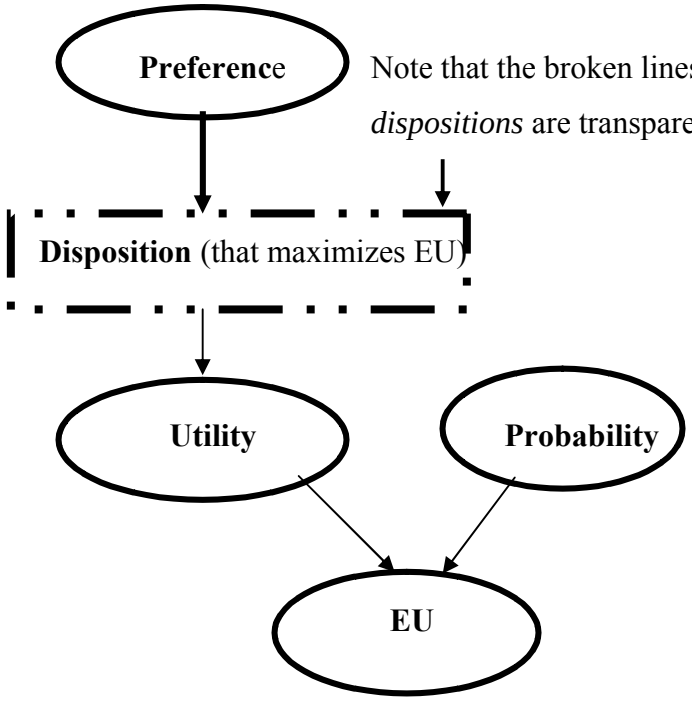
[142] Ibid, pp.185,186.

[143] MRTC is an essential element of the general argument that Gauthier advances that morals are by agreement. This is because MTRC associates rationality with utility-maximization at the level of dispositions. MTRC advances the argument for the rationality of a CM strategy, that is to say, the rationality of constraints in strategic or cooperative situations. For this reason, rather than referring to MTRC, in contrast to TRC, I shall simply refer to it within the broader standpoint of Gauthier's moral contractarianism, namely Mb(CM)A. So, except where I indicate otherwise, when I refer to Mb(CM)A I should be understood as referring to MTRC.

choice contexts, they are central to the multi-tracked framework for solutions in the test of application of the problem of secession.

In addition to replacing (a) and (b) with (a^1) and (b^1), Gauthier introduces a third element or property to rationality. Let us call this third element (c) the property of *translucency*. This property claims that in cooperative situations, an agent's true intentions or dispositions are not unknown. If we redraw figure 3.2.2b so as to reflect this additional element to rationality, we would have the following, as figure 3.2.2.1c illustrates.

Figure 3.1.2.1c: **Translucency of Dispositions and Constrained Maximization**



Decision-makers, Gauthier claims, have a better than equal chance of correctly identifying the characters of others. He writes, "We…[assume] that persons are neither transparent nor opaque, so that their disposition to cooperate or

not may be ascertained by others, not with certainty, but as more than mere guesswork."[144] A fully transparent agent is one whose dispositions are completely visible to others. A fully opaque person is one whose dispositions are completely invisible to others. And a fully translucent agent is one whose dispositions are sufficiently visible to others. No one claims that humans are fully transparent, but the observation that we can often successfully predict what others will do suggests that we are not fully opaque but are at least translucent.

The trustworthiness of our current evaluations of the morality of those we are dealing with depends, if we might say, on the difficulty of lying convincingly about one's intentions over an extended period of time or in face-to-face interaction. It seems that there are subtle clues of dishonesty and honesty that most of us are capable of detecting and picking up. We might call those "instinctive hunches" about who is and is not trustworthy, hunches that seem accurate more often than not. If this is so, then we can say that the probability of identifying people's dispositions or character is between 0 – 1. If 1 is the probability that others are transparent, then the probabilities that others are opaque and translucent, we might say, are 0 – 0.5 and 0.6 – 0.9 respectively.

Gauthier's modification of TRC seems to fit with the sorts of choice contexts that cooperative activities represent. The defining characteristic of cooperative activities is that they are generally strategic. Strategic choices or contexts contrast with parametric contexts in a relevant sense. In parametric contexts one's choices do not affect the choices of others, that is an individual's choice is independent of the choices of others, whereas in strategic contexts an

[144] MbA, p.174.

individual's choices is partly dependent on that individual's expectations of the choices of others, and vice versa. The disposition to SM, namely, the strategy to act exclusively and directly on utility-maximizing actions, is rational from a parametric context because one's action is "the sole variable in a fixed environmental." To say this is to say that at the level of each individual choice it may make sense to be an SMer.

A helpful illustration is this: suppose I am contemplating seeing a weekend baseball game. Suppose that my options are (i) the game between Toronto Blue Jays and the Texas Rangers in Toronto, and (ii) the game between the Boston Red Sox and the New York Yankees in Boston. Because this is a parametric situation, what should matter to me as a rational person who is seeking to maximize expected utility is to go to the game that maximizes EU. Suppose finally that the EU of (i) is 12 and the EU of (ii) is 5. Since I am rational and since I seek to maximize expected utility what is rational for me to do here, according to TRC, is for me to go to the game in Boston.

Gauthier agrees with TRC that it is rational for me to go to the game in Boston. In a parametric context, it would be rational to be an SMer and maximize expected utility. However, Gauthier argues, that the SM strategy is neither utility-maximizing nor rational in strategic contexts because one's choices are closely linked to the expectations and the choices of others. The babysitting example is an example of strategic context, as is any PD. Given that one's behavior is but one variable among others in strategic contexts, such that one's choice is responsive to one's expectation of others' choices, in the same way that their choices are

responsive to their expectations of one's choices, one maximizes expected utility by strictly adopting the disposition to CM. By focusing on directly utility-maximizing actions, TRC blurs the distinction between strategic and parametric contexts and thus mistakenly recommends the same strategy or behavior in both contexts.[145]

3.1.2.2 CM and the PD

As I noted at the beginning of this section, CM is an important component in Mb(CM)A. Gauthier calls it "the most fruitful idea in *Morals by Agreement*."[146] How fruitful is it? In other words, how does it fare with the general problem of rational compliance, which arguably is essentially the PD? In particular, how does CM deal with the non-iterated PD, for which cooperation is a one-time event? Since Mb(CM)A presents itself as a rational morality; a morality of constraints and, I must point out, of hypothetical agreement, the risky steps for it in moving from hypothetical agreement to actual moral constraints or to "actual moral practice"[147] is that it must solve the problem of the apparent rationality of being a freerider on cooperative behavior of others. The PD, I noted in chapter one, arises when the equilibrium outcome diverges from the optimal cooperative outcome, with the equilibrium outcome being the noncooperative action and the optimal outcome the cooperative action.[148]

---

[145] Ibid, p.21.
[146] Gauthier, "Uniting Separate Persons," p.185.
[147] MbA, p.168.
[148] The study by Eldar Shafir and Amos Tversky on choice under uncertainty provides interesting ways of looking at cases of the PD. The study shows that there are more factors to consider when evaluating what strategy people adopt in the PD. The trial subjects who were presented with PD games displayed on a computer screen one at a time had to, on each trial, choose whether to

Gauthier agrees with the general characterization of the compliance problem as instantiated in the PD. However, he disagrees that the unique solution in the PD, iterated or non-iterated is defection, i.e. Pareto-suboptimality. Part of the fruitfulness of CM as a strategy, Gauthier argues, is that it maneuvers agents away from suboptimal outcomes to optimal outcomes. Gauthier's argument here rests on the view that under reasonable and plausible circumstances CMers having internalized principles that govern their choices and actions generally do better than those that are solely motivated to choose directly utility-maximizing actions. So, for Gauthier, even when the optimal outcome does not converge or coincide with the equilibrium or suboptimal outcome the optimal outcome is still the preferable outcome and a CM strategy allows us to achieve the optimal outcome instead of the equilibrium outcome.

So, consider the following example of two persons contemplating cooperation, a familiar example employed in the literature to tease out both iterated and non-iterated PD.[149] Mabel and Abel decide to cooperate, but this time not as

---

compete or cooperate by pressing the appropriate button. When told that the other has elected to compete, the great majority of subjects reciprocated by competing, only 3% cooperated. When informed that the other has chosen to cooperate, 16% of subjects choose cooperation. When they were not told what the other's strategy is, a larger percentage (37%) elected to cooperate. This suggests, among other things, (1) that cooperation goes up when subjects know their opponent's have elected to cooperate, and (2) subjects are more likely to cooperate when they are ignorant of the other's strategy *possibly* in the hope that they expect the other to cooperate, or that they believe that the other, like them, is similarly disposed to cooperation. See Eldar Shafir and Amos Tversky, "Thinking through Uncertainty: Nonconsequential Reasoning and Choice in *Preference, Belief, and Similarity*, Eldar Shafir (ed.), Cambridge, the MIT Press, 2004, pp.701-727.

[149] The example of a two-person interaction is quite common in the literature on rational choice, partly because as a paradigmatic example of cooperation it brings out the salient aspects of cooperation and the compliance problem. See Holly Smith's example of the two fishermen, "Deriving Morality from Rationality" in *Contractarianism and Rational Choice*, pp. 229-231. Also, Robert Sugden's example of Alice and Bruce, in "The Contractarian Enterprise" in *Rationality, Justice and the Social Contract: Themes from* ***Morals by Agreement***, pp. 18,19. As well, see Gauthier's example of Jones and Smith, in "Why Contractarianism?" in *Contractarianism and Rational Choice*, pp. 24, 25.

money market investors but as farmers. Their farms, which are adjacent to each other, are situated two kilometers from a mountain. In their yearly migratory walk mountain goats visit the farms twice a year, each time from either side (west and east sides) of both farms. Each visit destroys a large portion of their crops, the cost of which is $500 for each farm. Mabel (or Abel) can individually choose to build a wall on her side of the farm and this would prevent the goats from getting into the farms. The cost of erecting and maintaining a wall per year is $900. If only one of Abel or Mabel builds, the visit would be reduced to one per year. In this circumstance, the one that builds incurs an annual cost of $1400 ($500 for one goat visit + $900 for building). The one that refused to build pays a yearly cost of $500 ($500 for one goat visit + $0 for not building). If both build, their cost annually is $900 each ($0 for no goat visit + $900 for building). And if neither of them builds, they each incur a yearly cost of $1000 ($1000 for two goat visits + $0 for not building).[150] The dilemma posed by this situation can be represented in the following standard PD matrix.

In the classic form of the PD (like the one involving Abel and Mabel) pursuing individual interests and welfare does yield a suboptimal outcome. However, both benefit and achieve an optimal outcome if they adopt a joint strategy whereby they agree to build a wall on both sides of their farms. One way to achieve this is to constrain their behavior by principles prohibiting purely selfish

[150] Mabel and Abel's situation can be formulated in the form of the PD precisely because it offers (a) suboptimal and optimal outcomes, and (b) different options as well as utilities, and the possibility of defection—choosing a different option from the one that is jointly optimal, or has been agreed on.

behavior, i.e. by a principle requiring both of them to build. Such a constraining principle would qualify in this sense as moral principles or constraints.

Figure 3.1.2.2: **The Prisoner Dilemma with Matrix Showing Money ($)**

		Mabel	
		Build	Don't Build
Abel	Build (Cooperate)	–900, –900	–1400, –500
	Don't Build	–500, –1400	–1000,–1000

But suppose that the weather situation in the mountain hampers simultaneous building by Mabel and Abel. If both walls can only be erected consecutively, what stops either Mabel or Abel from defecting from the agreement (refusing to build) after the other has built, since adopting this strategy cuts the total cost to $500, less than what one loses if one were to cooperate by building (i.e. $900)? Otherwise stated, why is it rational for Mabel to keep her commitment to the agreement she made with Abel, even when she apparently benefits from defecting? This is similar to the babysitting example that I discussed in chapter one. Given that refusing to babysit for you gives me greater expected utility, why should I babysit for you this weekend after you babysat for me last weekend?

In classic form of the game, TRC says it is rational for actors to defect because cooperating is always strictly dominated by defecting. In iterated versions of the game, when the game is played repeatedly and the actors know in advance

the number of steps, the rational choice for each actor, according to TRC, is to defect repeatedly. In non-iterated versions of the game, when cooperation is a one-shot event, TRC says the rational choice for each actor is defection. The only time it is rational, according to TRC, for each actor to cooperate is when the game is iterated infinitely, that is, when it is played endlessly or for a random number of times.

If Mabel were to adopt a TRC's view of rationality (which is fundamentally an SM view of rationality) she would perform the dominant action. Thus, she would ignore whatever Abel does or choose to do and not build. But we have seen that a TRC's view of rationality is mistaken. The mistake, we saw, arises from the identification of rationality with directly utility-maximizing actions and the TRC's failure to decouple parametric from strategic situations. Abel and Mabel's situation is a strategic one. Abel's choice and expectation are dependent upon and responsive to the choice and expectation of Mabel and both parties benefit if they build. Mabel reasonably expects to benefit more from promising to build than from any other alternative strategy since the only way she gains Abel's agreement is by promising to build. After Abel has built, although she has no directly utility-based reason to build, but because she expects to be doing better honoring her promise than she would have done had she not promised to build, she has sufficient reason to build.

By not distinguishing between choice in strategic contexts from choice in parametric contexts and by identifying rationality with directly utility-maximizing actions, TRC defines rationality narrowly. If a choice or action is rational if and

only if it is directly utility-maximizing, then one is permitted to break commitments when they are not directly utility-maximizing. But to restrict promise keeping to when it is directly utility-maximizing is to reduce the benefits available to agents. This is because, as Gauthier rightly puts it, "among persons *whose rationality is common knowledge*, only promises that require limited compliance will be made. *And opportunity for mutual advantage will be forgone*."[151]

A defender of TRC might agree with Gauthier that the opportunity for mutual advantage is a significant motivation for agents to cooperate, especially when there is the possibility of future interactions. In iterated (especially infinite or random) situations, there are mechanisms—such as reputation, induced reciprocity, punishment—that affect choices, and agents that are committed to keeping promises only when they are directly utility-maximizing would be excluded from cooperation. They may seem to benefit in the short run, but in the long run, because they would be detected and branded as defectors and back-stabbers, they would be punished and excluded from further cooperative ventures.

In non-iterated situations, as is the case with Mabel and Abel, the mechanism of reputation and punishment does not seem so attractive.[152] Their interaction is a one-time event, and with no opportunity for further interaction,

[151] Gauthier, "Why Contractarianism?" in *Contractarianism and Rational Choice*, p.24, emphases are mine.

[152] We might even assume that after Abel builds, Mabel decides to sell her farm and move away. The person she sells her farm to is not interested in farming. In which case, it might be argued that considerations of reputation and the expectation of future interactions and benefits would not play much role in the choice of Mabel not to build after Abel has built. We might suppose too that both Mabel and Abel are aware that the other person is selling his or her farm at the end of the month, and after selling would be moving far away. Both situations are clear cases of non-iterated PD, which is usually taken to be the most problematic for a contractarian account of rational compliance. Given that Gauthier admits that CM does not appeal to the idea of reciprocity and since he thinks that "constrained maximizers may cooperate even if neither expects her choice to affect future situations" (MbA, fn.170), he needs to show how in the case of Mabel and Abel it is rational for either to cooperate, even if both expect not to interact with each other after this one interaction.

Mabel's defection would seem beneficial. The bottom line is for an agent (Mabel) to cheat and defect in non-iterated PD situations because it is directly utility-maximizing, but to cooperate and build in infinitely or randomly iterated PD situations because it is not strictly speaking directly utility-maximizing. For if the situation is iterated, given that an agent is fully rational she would cooperate if she estimates that she would be caught and punished, but in the absence of punishment she does well by defecting.

However, if we consider what Gilbert Harman says about the relationship between an individual's past reputation and the choice of whether or not to cooperate, then it is right to say that in many PD situations, punishment and induced reciprocity are not necessary to encourage cooperation. Harman makes the point that prior cooperation by individuals plays a role and contributes to identifying likely cooperators.[153] If we are sufficiently able to tell who is a likely cooperator either from their history of past interactions or from their dispositions, then unlikely cooperators would be branded as such and would be excluded from cooperation, both those that are iterated and those that are one-time. And because they would be excluded from this one-time cooperation they would lose out on the benefits of mutual advantage.

The crucial point here is that the dispositions of agents are not veiled. The property of translucency, which Gauthier introduced into rationality, can be taken to lie between 0.6 – 0.9, such that in PD or other situations agents are sufficiently able to predict the dispositions of other agents. Since property *c* lies between 0.6 –

[153] Gilbert Harman, "Rationality and Agreement: A Commentary on Gauthier's *Morals by Agreement*," in *Social Philosophy and Policy*, vol. 5, 1988, p.5.

0.9, we suppose that Abel is adequately able to predict what option Mabel would choose. Given that he is able to do this, he would tailor his choice to that of Mabel. If Mabel has adopted a disposition to SM, Abel would know this. Since she has formed a disposition to SM, she would not build and if she would not build, Abel would not to be disposed to build. This, however, leaves both of them worse off since both lose out on the benefits from cooperation. But given that they are both rational and desire to maximize their outcomes and since they both can tell what disposition the other person has formed, each would adopt the disposition to build.

To summarize, Abel and Mabel are both locked into a PD, a situation of choice under strategic context. Since this is a strategic context, Abel's choice to build is partly dependent on his expectations of the choice of Mabel to build. Given the ability of Abel to predict the true intentions of Mabel, he effectively models his choice on what he expects her to do.[154] If Mabel is disposed not to build, Abel will form the disposition not build. If Mabel forms the disposition to build, Abel will adopt a similar disposition to build.[155] If he chooses the disposition not to build, he

[154] There seems to a connection between what the property of translucency assumes and the result of the study by Eldar Shafir and Amos Tversy. A variation of this experiment reveals a strong desire and behavioral pattern in subjects to model their choice on their opponent's decisions. Subjects were offered the opportunity to learn the opponent's decision before making their own choice if they pay a small fee. On 81% of the trials, subjects first choose to pay a small fee to discover the other's decision. This behavior may in fact suggest two things: first, that most people think it is important to put themselves in a position where they can access the behavioral pattern, or garner information about the strategy of their opponents, and second, that they consider such knowledge crucial to their own decision whether to compete or cooperate. This willingness to pay for information about the decision of others places some weight on both the reasoning regarding the interest of agents in predicting the dispositions of others and the agent's choosing of the disposition to cooperate provided others are similarly disposed. This may in fact be descriptive of the behavior of people, but it might very well be normative in the sense that it stipulates that it is rational to expend one's resources in ways that places one in a position to predict sufficiently well the dispositions of others. See Eldar Shafir and Amos Tversky, "Thinking through Uncertainty: Nonconsequential Reasoning and Choice," in *Preference, Belief, and Similarity*, pp.701-727.

[155] One way of looking at Abel's action is by comparing it to the great majority of the trial subjects, in the study by Eldar Shafir and Amos Tversy, who reciprocated the choice of the others to compete by competing, and the 16% who decided to cooperate because their partners choose to cooperate.

chooses the disposition because Mabel has adopted the disposition not to build. If Abel chooses the disposition to build, he chooses the disposition because Mabel has the disposition to build as well, and if Mabel has formed the disposition to build, she chooses the disposition because she knows the disposition provides her optimal outcomes and because she expects Abel to adopted a similar disposition to build.

Since both Abel and Mabel are able to read sufficiently the dispositions of the other and since they are rational and desire to maximize their outcomes, they will each form the disposition to build, and accordingly, Mabel will build after Abel has built. Thus considered, what provides Mabel the reasons to form the disposition to build and to actually build, after Abel has built, is the fact that Abel had formed a similar disposition. And once Abel builds, she has a rational motivation to build as well, because she would not have benefited (from Abel's action) where she not disposed to build, and Abel would not have built were he not expecting that she was disposed to build and would build when the time comes. This way of putting it thus suggests that the two crucial considerations leading Mabel to cooperate are the opportunities for utility-maximization and the fact that both she and Abel are able to predict sufficiently the dispositions of the other person.

A critic or an advocate of TRC may want to agree with Gauthier that Abel's ability to predict sufficiently Mabel's disposition and Mabel's ability to predict sufficiently Abel's disposition significantly adds weight to their consideration and choice to form the disposition to build and to build when the time comes. The critic

however, may point out that the opportunities for utility maximization and the ability to predict sufficiently the dispositions of the other are not enough to lead either to cooperate. How might the critic argue her point? The critic might remind us that each person's ability to predict the other's disposition is not 100%. Our dispositions, according to Gauthier are not *transparent* but *translucent*, namely, they lie between 0.6 – 0.9. If dispositions are not transparent, then it is possible for Abel and Mabel to mistake the other party's disposition and believe that the person is disposed to cooperate when in fact the person is disposed to defect or to not cooperate.

Besides the problem of mistaken dispositions, the critic could argue that the adoption of a particular disposition does not guarantee that individuals will act on the disposition they form. The point is this: that Mabel has a particular disposition, say, a CM disposition or a disposition to build is not an assurance that she would in fact build, when the time comes. It seems common for people to make commitments at time t_1 and not carry them out at time t_2. There is nothing in the disposition of Mabel that *guarantees* she would build when the time comes even if she were to reveal her disposition and commitment to build to Abel. So as the argument goes, the opportunities for utility-maximization and the ability to predict sufficiently the dispositions of the other person may add weight to Abel's and Mabel's decision whether either ought to cooperate or not, but they are not sufficient to lead each of them to cooperate or to refuse to build after the other person has built.

Gauthier might respond to these worries in two ways. Firstly, he might argue that because rationality and the dynamics of human activities favor cooperation, dispositions would be sufficiently uncloaked to prevent people from being mistaken about them. No one wants to be left out from the largesse of cooperation because he or she has not sufficiently convinced others that he or she is not a cheat and defector. Secondly, he could argue that the relationship between dispositions and actions is not as loose as the critic is suggesting. This argument requires providing a deeper analysis of rational dispositions. Earlier on, Gauthier has suggested that complying with the terms of a contract expresses an agent's rational disposition. That is, the disposition to cooperation is formed because of the agent's belief that cooperation is mutually beneficial for everyone. The disposition is not formed because one intends to cheat or defect. Rather, the manner in which the disposition is formed makes it rational for one to follow through on one's commitments even if on occasion compliance fails to provide greater benefits or maximize expected utility of those particular actions.

By arguing that dispositions commit us to actions, Gauthier is arguing that our dispositions forge an invariant connection to certain actions and by performing an action we disclose publicly a particular disposition, namely, a disposition that had previously been chosen because of our estimation of the benefits that such a disposition provide. Dispositions thus understood are rationally chosen. Dispositions are rationally chosen when we expect such dispositions to contribute to one's overall life. Rationality forbids us from having irrationally formed dispositions or "disposition disorder," namely, indistinct or fuzzy dispositions.

Rationally formed dispositions are better than irrationally formed dispositions not only because they are constitutive of rationality but also because they are essential to our life-plan. Since dispositions relate to an individual's character, they "lay that individual bare" for what and who the individual is.

To understand why it would be rational for Mabel to adopt a CM disposition rather than an SM disposition, Gauthier directs us to consider the way in which dispositions and commitments relate to actions. As a CMer, Mabel adopts a disposition to build conditional on Abel's having a similar disposition. For if it is possible for Abel to tell what dispositions (CM or SM) that Mabel has he would be in a position to reasonably guess whether or not she will keep her commitment to build. Hence, Mabel will reason in the following manner: "Abel would be irrational if he goes ahead to build even though he believes that I would not build. If Abel builds, he builds because he believed (rightly) that I was disposed to keep my commitment to build as well. Since people's dispositions are translucent, if I had not really been disposed to keep my commitment Abel would probably not have formed that belief. Thus, it is to my overall advantage that I not only have a CM disposition but also act on it. So I shall act on my CM disposition by carrying out my commitment, and so I will build."

Following David Copp, we can express Gauthier's argument about dispositions and actions in this way, "[An] agent's action expresses a disposition to do A in circumstances C if and only if the agent is disposed to do A in circumstances C, and the action is a case of doing A in circumstances C, and the

agent's disposition explains his action."[156] The relationship of disposition to action can be formulated as backward-looking and forward-looking, namely, moving from an action to a disposition, on the one hand, and moving from a disposition to an action, on the other hand. We move from an action to a disposition when we focus on a particular action explaining a person's disposition, and we move from a disposition to an action when we focus on a disposition pointing towards a particular action.

There is something eerily omniscient here; presumably, an omniscient being (of the sort in Newcomb's Problem)[157] looking down at agents engaged in various activities via their dispositions. By simply looking at their dispositions the being is able to tell accurately what their next actions would be. Rational actors need not be omniscient, but in expending themselves so as identify the dispositions of others, they can reasonably tell what actions fellow rational actors would perform in n^{th} context. This is a variant of Kant's analysis of maxims. If one can precisely formulate the maxim of a rational agent—maxims that implicate the agent as a rational "willer" or member of the kingdom of ends—one can easily tell what that agent's deepest motives or intentions are and ultimately what actions would be performed in n^{th} contexts *if that agent is consistently rational*.

We need to keep in mind that Gauthier's aim is to justify particular actions by appealing to agents' dispositions and then to justify the dispositions by appealing to their contribution to an agent's utility profile. The idea that actions

[156] David Copp, "Contractarianism and Moral Skepticism," in *Contractarianism and Rational Choice*, p.200 fn.
[157] I discuss Newcomb's Problem in chapter five, not in connection with the relationship between disposition and action, but in connection with DV.

express particular dispositions is a fruitful one for Gauthier. Indeed, the expressive character of actions *qua* dispositions provides Gauthier with part of what he needs in arguing for the rationality of acting on moral reasons. After all, one's dispositions are part of one's character and they disclose, expose or lay bare the sort of person one is. Furthermore, this approach gives Gauthier some of what he needs to advance his claim that the possession of a CM disposition and the ability to sufficiently predict what dispositions people have, considered along with the opportunities for utility-maximization, are sufficient to lead us to cooperate even when the action in question does not maximize expected utility.

What makes the above line of reasoning appealing is that it allows us to be able to say which dispositions are good or rational and which are not. Once we identify which dispositions are rational or good we can tell which actions are rational or good since an action can only be rational if the disposition it expresses is itself rational. We might say that a disposition is rational or good if that disposition positively promotes an individual's life-plan. Or stated differently, dispositions *qua* actions that combine to frustrate a person's end-goal are not rational or good and those that promote that person's end-goal are rational or good. Since one's life-plan is tied to one's utility profile and since a CM disposition, according to Gauthier, enhances the condition for the possibility of cooperation and mutual advantage, one promotes one's life-plan when one adopts those dispositions that encourage cooperation.

Holly Smith and Geoffrey Sayre-McCord have individually suggested that for the above argument to be successful, Gauthier needs some principle that

connects dispositions to actions. Whereas Smith calls the connecting principle the "rationality of perseverance principle,"[158] Sayre-McCord calls it the "transitivity of rationality principle."[159] Either principle ensures that the rationality of dispositions carry over to the actions. That is, the rationality of disposition transfers or carries over to the manifestations of the dispositions, i.e. to those actions.

Gauthier's argument, in MbA and elsewhere, that disposing oneself to comply "strictly" requires compliance seems to indicate that he subscribes to either principle. In "Afterthoughts" Gauthier writes, "If it is rational for me to adopt an intention to do *x* in circumstances *c,* and if *c* comes about … then it is rational for me to carry out *x.*"[160] So if I have disposed or committed myself to build at time t_1, it would be rational, Gauthier says, for me to actually build at time t_2 when the occasion arises even though other payoffs (defecting, noncooperation, cheating, or adopting an SM or other strategy that aims strictly at utility-maximizing actions) exist and seem attractive.

Note that Gauthier not only plays on the idea that the rationality of dispositions carries over to actions but also claims that having a disposition to do *x* at time t_1 does in fact "lead" to performance at time t_2. To bring out this line of reasoning, we have to state the two ways dispositions can lead to actions. The first version, "the strong thesis," is what Smith calls "the *causal efficacy thesis*: the

---

[158] In "Deriving Morality from Rationality," *Contractarianism and Rational Choice*, Smith defines the *rationality of perseverance principle* (RPP) this way: "If it is rational for an agent to form the intention to do A, then it is rational for the agent to actually do A when the time comes (assuming the agent acquires no new information and has not altered her values)," p. 244.

[159] Sayre-McCord, "Deception and Reasons to be Moral," in *Contractarianism and Rational Choice*, p.183.

[160] Page 159 quoted in Smith (p.244); see also MbA, p.186.

thesis that forming an intention to do A will cause the performance of A."[161] The second, which I call "the weak thesis" is *the probability efficacy thesis*: the thesis that forming a disposition will probably or very likely lead to performing some particular action. The causal efficacy thesis assumes a strong kind of behavioral determinacy while the probability efficacy thesis assumes a weak kind of behavioral determinacy. Evidently, there are problems associated with this way of relating dispositions with actions. We briefly examine some of them below.

3.1.2.3 Five Problems for CM

Problem 1

Certainly, Gauthier's strategy commits him to the causal efficacy thesis if he is to argue that X's choosing CM will cause X to act on the chosen disposition and so "induce [Y] to build and so maximize [X's] utility."[162] But, as the objection goes, the causal efficacy thesis is an implausible assumption. Smith states what he takes to be the problem with the assumption:

> I find it quite implausible to assume that any intention of mine *inevitably* causes my subsequent carrying out of that intention: some do, but some do not. Upgrading the *kind* of mental state I form (to a commitment or resolution) does not change this fact. Of course we often change our minds when we acquire new information, or when we adopt new values… it is implausible to suppose our commitments always compel our future

[161] Smith, p.235.
[162] Smith, "Deriving Morality from Rationality," *Contractarianism and Rational Choice,* p.235.

acts—especially in the kind of case in question, where considerations of utility press the agent to change her mind when the times comes.[163]

Response

I agree with Smith that Gauthier's strategy seems to require the causal efficacy thesis. It is unclear however, why she thinks it is an implausible assumption. Note that the latter part of Smith's criticism is similar to the criticism that the critic raised earlier, i.e. the criticism that people may commit themselves at time t_1 to do *x* but fail to do *x* at time t_2. Like the critic's objection, Smith's objection seems to ignore a crucial aspect about dispositions and commitments and their overall contribution to an individual's life-plan. Recall the earlier point I made about dispositions, the point that dispositions partly define an individual and that rationally chosen dispositions are those that are chosen because the individual estimates that they contribute favorably to his or her life-plans. Consequently, it would seem irrational for an individual not to choose those actions that express that individual's particular dispositions insofar as those dispositions *qua* actions contribute favorably to his or her life-plans.

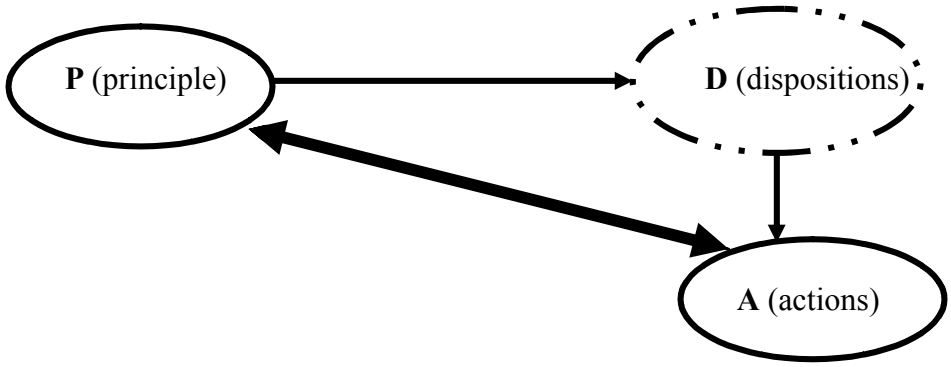
Could it be said also that dispositions express principles and are chosen because they reflect, to some degree, a person's character or the kind of person one is, in addition, of course, to the fact that they foster conditions that leads to optimal outcomes, namely, EU? If we suppose this is the case, then the claim would be that when a person disposes himself or herself or commits to act in certain ways, say to cooperation, at time t_1, besides of course the connection of dispositions to EU, he

[163] Ibid, p. 236.

or she is notifying others that this is who I am. On this interpretation, dispositions are not just constitutive of a person's life-plan but form part of the critical elements of that person's character or identify.

Acting in certain ways as Nozick has argued expresses some aspect of one's character, i.e. it reveals the sorts of principles one holds.[164] The point is that dispositions play some expressive role, that is when we choose a particular disposition we are partly using it to express some kind of principle that we hold about ourselves. As objects of rational reflection, a disposition expresses or points towards a principle and such a disposition is expressed by an action, in the sense that I choose an action because of the disposition that it expresses. This relationship can be formulated as principle (P) = disposition (D), D = action (A), therefore P = A, as illustrated in figure 3.2.2.3.

Figure 3.1.2.3: Transitive Relation amongst Principle, Disposition and Action



The tick double arrow pointing from **A** to **P** and from **P** to **A** indicates that the transitivity moves from **P** to **A** because **A** expresses **P**, and from **A** to **P** because **P** is disclosed by **A**.

[164] Nozick, *The Nature of Rationality*, ch. 1 "How to Do Things With Principles," pp.3-40.

I choose a disposition because it expresses a particular principle and my having the disposition is expressed by the action that I choose. Let us suppose that I define myself as a just and fair person and so have as my principle "the principle of fairness." On this view, the dispositions I choose would reflect this principle, namely, I would be disposed to engage in just and fair practices. To say that I am disposed to engage in just and fair practices is a shorthand way of saying that I am disposed to choose actions that express "the principle of fairness." Some kind of transitivity is at work here, for if P is expressed by D and if D is expressed by A, then P is expressed by A. That is, P = D, D = A, therefore P = A. The sort of semantic representation and transitivity I am suggesting need not be fine-grained or robust. Gautier's argument that dispositions commit one to actions would be boosted if it can be shown that some transitive relationship of the sort I sketched above holds.

In specifying this kind of transitivity and as is common with all forms of transitivity, we can simply jump over from P = D and D = A to P = A, as the tick arrow in figure 3.2.2.3 shows. We might imagine someone walking her dog at noon everyday and expect that she would tell us that the reason she walks her dog everyday is that she is acting on some principle.[165] If we believe that there is something deeper going one we might ask her the following question, "what principle are you acting on when you walk your dog every day at noon for seven years?" Her response that she walks her dog at noon every because she believes it

[165] Or as the story goes about Kant leaving his house and walking through the same path the same time for most of his adult life. Kant would certainly have an explanation for this, an explanation that isn't random or arbitrary but one that embodies the principle he acts upon, maybe the "principle of consistency or reliability" or some other principle.

is good for both her health and that of the dog may likely not satisfy us. "But that can't be all," may likely be our respond to her answer.

If we believe that there is something deeper going one then we would expect her to tell us that the reason she walks her dog every day is that she is acting on some principle. Even though she does not tell us this, we might be able to link her belief about the relationship between her health, the dog's health, and walking the dog every day to some principle. We might connect the belief that a healthy lifestyle is good to some principle, say, the "principle of living a healthy lifestyle." This principle, we might say, determines the actions she chooses, including the act of walking her dog regularly. The idea here, as Nozick writes is that:

> Principles constitute a form of binding: We bind ourselves to act as the principles mandate. Others can depend upon our behavior, and we too can benefit from others' so depending, for the actions they thereby become willing to undertake can facilitate our social ease and interactions, and our own personal projects as well.[166]

Principles, as Nozick seem to suggest above, are important in many ways. They tell us what kind of people we are; they define our identity in some very robust ways, so to speak. "I am a person with *these* principles."[167] Besides, even were the future to bring the person with principles some inducement to deviate from them, "we can trust that he will not, and we can rely upon this in planning and

[166] Nozick, *The Nature of Rationality*, p.10.
[167] Ibid, p.12.

executing our own actions."[168] Principles thus, serve as rich source of information about people. They tell us at time t_1 what actions people *would do* at time t_2.

Nozick, as does Gauthier, clearly thinks that following through on commitments or principles contributes to an individual's utility profile. However, he goes beyond this by arguing that following through on commitments or principles cement social relationships—that is they help in building trust and facilitating social ease and interactions—which could again add to an individual's utility profile, a view that Gauthier would be willing to endorse. Imagine a society where everyone breaks every single commitment. People would hardly be able to plan or accomplish anything. Everyone would have to live *only* for the *moment*. Might it be said that we are socialized to value keeping commitments and acting on principles and on reflection we come to realize the importance they play in boosting our utility profile?

To be sure, acting at time t_2 on a commitment made at time t_1 may not be directly utility-maximizing, but we do recognize that overall it is beneficial. Isn't that the whole point about commitment? Isn't the essence of commitment to "keep us from the path," as Nozick puts it, "of inducement or temptation to deviate"[169] from agreements, i.e. the temptation to identify with directly utility-maximizing actions. Besides performing some interpersonal function in reassuring others that one will get past temptations, principles act as barriers to a person's following the desires or interests of the moment. This is similar to Ulysses binding himself to the

[168] Ibid, p.9.
[169] Ibid, p.9.

mast of his ship to prevent the desires and interests of the moment from plunging him to death.[170]

If Gauthier is right that we are bound by our principles, if he is right that our principles bind us to perform some particular acts because P = D, D = A, therefore P = A, then he would have shown in some way, *contra* Smith, that our commitments *always compel* our future acts—especially in the kind of case in question, where considerations of utility press the agent to change her mind when the times comes. Of course, the nature of the binding has to be specified. Is it like the type expressed in marriage vows, "till death do us part," or the kind of binding that is irreversible? Gauthier does seem to suggest the kind of binding that is irreversible when he discussed cases of failed threat and deterrent. Smith brings forth the following example, and Gauthier discusses a variant of it in MbA:

> Consider a case in which a telepathic burglar threatens to steal all my household valuables. I know that if I form the intention of blowing up the house with the burglar and myself inside, it is nearly certain that he will be deterred. According to the strong perseverance principle, it is now rational for me to blow up the house and kill myself, merely because I previously formed the intention of doing so under these circumstances. But no one, I think, would want to agree with this.[171]

---

[170] *Ulysses and the Sirens*. In Greek mythology, the Sirens' song was so charming and captivating that sailors steered their ships onto the rocks from which the Sirens sang in order to get closer to them and to hear their songs better. When Ulysses' ship was passing close to the Sirens, upon the advice of Sorceress Circe, he instructed his sailors to stop their ears with wax ears so that they could not hear the song, and also to bind him to the mast of his ship with his ears unblocked, so that he may hear the song and yet would be unable to plunge to his death in an attempt to get closer to the Sirens. Moreover, he instructed his sailors to ignore his orders to untie him (and to tie him more), orders which he knew he would give when he came close to the Sirens.

[171] Smith, "Deriving Morality from Rationality," Contractarianism and Rational Choice, pp. 246, 247.

Gauthier's response to cases of this nature is that they fall into the category of weakness or imperfection, a "second-best rationality."[172] This is because the rationality that I display when I formed the intention to blow up the house, the burglar and myself blurs the difference between satisficing or suboptimality and maximizing or optimality. Forming the intention to blow the house and actually blowing it is not optimal, hence, not fully rational. Gauthier writes,

> For although it may be rational for us to satisfice, it would not be rational for us to perform the action so chosen if, cost free, the maximizing action were revealed to us. And although it may be rational for us to adhere to principles as a guard against wish fulfillment, it would not be rational for us to do so if, beyond all doubt, the maximizing action is revealed to us.[173]

Given that Mb(CM)A rules out all kinds of irrationality, the sorts of agents that a rational morality is concerned with are sufficiently rational agents, agents who do not get hoodwinked into making irrational threats and agents that are capable of forming rational dispositions, namely, dispositions that positively promote their life-plans. And since "the entire point of disposing oneself to constraint is to adhere to it in the face of one's knowledge that one is not choosing the maximizing action,"[174] an agent that rationally disposes herself to threat enforcement or threat resistance is rationally required to carry "out a failed threat" or "resist despite the cost to herself."[175]

---

[172] MbA, p.186.
[173] Ibid, p.186.
[174] Ibid, p.186.
[175] Ibid, p.186.

To summarize, our disposition or commitment to comply with agreement and to cooperate with others, we may suppose, defines the kind of principle we act on. Being committed to do *x* is to dispose oneself to *actually* do *x*. This disposition, which contributes to our overall life-plan is a reflection of our deepest principle, i.e. the principle about the sort of person we are. And we act on this disposition when the appropriate occasion presents itself. However, given that Mb(CM)A advances an instrumental conception of rationality, the sorts of dispositions *qua* principles Gauthier is able to appeal to are those that are non-eternally binding.[176] For him to appeal to eternally binding principles he has to provide an argument that demonstrates that whatever dispositions an agent chooses, they are those, 'come what may,' that maximize that agent's EU. I am not sure if it is possible to provide such an argument, what will be accomplished by such an argument, nor what such an argument would look like.

Problem 2

Suppose we grant the following to Gauthier: one, that dispositions transfer to or carry over to actions, and two, that the causal efficacy thesis—forming an intention to do *x* will induce the performance of *x*—is plausible. How does this prove that dispositions are rational?

[176] Eternally binding principles would be those that discount all connections with utilities, principles that are non-instrumental, the sort that lead to acting on "till death do us part" even when it is apparent that in this situation, EU (or SU on a DV account) has been extinguished for one or both of "us." Kant's categorical imperative—acting for the sake of duty, or acting not based on heteronomous considerations—might be a good candidate for eternally binding principles. I will always tell the truth regardless of the instrumental value of the action. I will tell the 'inquiring murderer' where his victim is headed, notwithstanding what the outcomes are.

Response

Gauthier provides what appears to be a response to this problem in a fairly long passage:

> What underlies constrained maximization is the recognition that such a person would be less successful if she were to include among her reasons for acting exclusively and directly from her utilities, than if she were to include among her reasons other considerations only indirectly related to utilities. Among these considerations would be the execution of a plan or the honouring of a commitment even if in so doing, she would not, and would know that she would not, be maximizing her expected utility at that time, *given* (i) that the plan or commitment had been rationally undertaken (in terms of her expected utilities) and (ii) that in executing or honouring it she would expect to be doing better (in those terms) than she would be doing had she not undertaken it.[177]

Gauthier's argument for the rationality of dispositions can be outlined as follows:

> **P1**: A particular disposition is rational if and only if it in some way relates to one's EU i.e. makes one more successful or promotes one's life-plan.
>
> **P2**: The reason for adopting a particular disposition is that we rationally believe at time t_1 that the disposition maximizes one's utilities. As rational agents we have rationally evaluated the circumstances and come to the conclusion that having such and such dispositions maximizes our utilities.

[177] Gauthier, "Uniting separate Persons," in *Rationality, Justice and the Social Contract: Themes from **Morals by Agreement***, pp. 185, 186.

P3: Carrying out the disposition (commitment) at time t_2 only expresses and reinforces the earlier rational belief that the adoption of the disposition is rational.

P4: We form CM disposition at time t_1 rationally believing it to maximize our utilities.

Conclusion: Therefore, the CM disposition we adopt when we choose cooperative activities is rational.

Problem 3

This problem questions Gauthier's argument for limiting dispositions or strategies to SM and CM. Again, we turn to Smith's characterization of the problematic. Gauthier, Smith says, "mistakenly assumes both that (a) my only options are CM and SM, and (b) my partner's only options are CM and SM."[178] But there are strategies other than CM and SM that yield greater benefits and until Gauthier rules these out as irrational, he cannot justify CM as *the* rational strategy.[179]

Both Smith and Copp suggest different alternatives to CM. Although Smith offers "unconditional cooperation" and "radical cooperation"[180] as possible

[178] Smith, "Deriving Morality from Rationality," *Contractarianism and Rational Choice,* p. 238.
[179] Copp argues on page 221 in "Contractarianism and Moral Skepticism" in *Contractarianism and Rational Choice* that for CM to be a rational disposition, Gauthier, "would have to show that it is at least as productive of utility as *any* alternative disposition. As well, Peter Danielson argues that the dispositions of "reciprocal cooperation" and "counteradaptive cooperation" yield higher utilities than CM dispositions in two-person games and many-person games respectively. See his "Closing the Compliance Dilemma: How it's Rational to be Moral in a Lamarckian World," in *Contractarianism and Rational Choice*, pp. 291-322. See also Danielson's "The Visible Hand of Morality: Review of David Gauthier's, *Morals by Agreement*," in *Canadian Journal of Philosophy*, vol.18, no.2, June 1988, pp.373-383 where he argues for Reciprocal Cooperation as an alternative to CM.
[180] Unconditional Cooperation states that I do x, (i.e. build my wall) no matter what the other person does. Radical Cooperation states that I do x if and only if the other person has chosen to unconditionally cooperate. Reciprocal Cooperation is the strategy to cooperate "only when

replacement to CM, she acknowledges that they may not necessarily provide greater utilities in strategic contexts. On his part, Copp thinks that "reserved maximization" yields greater utilities than CM. A reserved maximizer, he says,

> has exactly the disposition of a constrained maximizer, except that he will violate a requirement of cooperative scheme whenever he has the opportunity to win a jackpot. He will take opportunities to make very great gains in utility, when the probability of detection is very low. For example, unlike a constrained maximizer, he may steal the money from a lost wallet, provided enough money is involved and provided he is quite sure he was not observed finding the wallet. A person might do better as a reserved maximizer than a constrained maximizer.[181]

Response

What will be Gauthier's response to this problem? Given that the aim of forming dispositions and adopting a particular strategy, according to Gauthier, is to encourage cooperation and to maximize one's utilities, no person will form the disposition to be an unconditional cooperator. A disposition to "unconditional cooperation" like "broad complaint" is an open invitation for exploitation and freeriding as "others will have every reason to maximize their utilities" at that person's expense, by offering cooperation on terms that offer her but little more than she would expect from noncooperation.[182] Furthermore, this disposition fails the demand of mutual advantage, a necessary condition for human interactions. It is a disposition, as Gauthier rightly notes "inimical, not only to its own survival, but

---

necessary, that is, with and only with those who cooperate with and only with cooperators," Danielson, "The Visible Hand of Morality," p.377.

[181] Copp, "Contractarianism and Moral Skepticism," p. 221.

[182] MbA, p.178.

to that of any form of cooperation. For a world of unconditional cooperators is easily invaded by straightforward maximizers."[183]

The larger plausibility of "reciprocal cooperation" depends on the condition of *transparency*. In the absence of realizing this in cooperative acts, which involves rational persons, who are only *sufficiently translucent*, reciprocal cooperation turns out not to be a superior or better strategy. In any case, a reciprocal cooperator does not reduce each individual cooperative action to the expected behavior of others in such a way that she expects a net gain from the action that is performed. To this extent, and insofar as appropriate feelings or the affective capacity for (rational) morality is integral to cooperation, reciprocal cooperation turns out to presuppose a key aspect of any disposition that is incorporated into CM. I examine the relationship between the affective capacity for (rational) morality and cooperation in the last section of this chapter.

Gauthier's modification of the rationality of the received view of rational choice theory seems to rule out "radical cooperation" and "reserved maximization" as rational strategies. Radical cooperation seems undercut by the causal efficacy thesis. For, if my disposition to build will induce my building when the time comes, then this will sufficiently guarantee performance for anyone who might opt for radical cooperation. In any event, even if what a radical cooperator is looking for is not some kind of guarantee, it is unlikely she would maximize her utilities given that she only cooperates with unconditional cooperators. Since unconditional cooperators recognize that a disposition to unconditional cooperation is inimical to

[183] Gauthier, "Moral Artifice: A Reply by Gauthier," p.401.

its own survival and any form of cooperation, the argument that one can exploit them seems quite moot.

As a strategy, reserved maximization reminds us of the core problem of rational compliance: I ought to maximize my utility profile at the expense of others if I have the opportunity to do so. Reserved maximization relies on the assumption that detection is low and that being opaque is costless. But this seems wrong. It is important for agents to be able to predict what others will do in strategic contexts and frustrating others by masking one's dispositions does not help a person disposed to reserved maximization. In cases where people find it difficult to recognize the dispositions of others because they have masked them, they are more likely to be avoided, if not completely excluded from interactions. At any rate, since Gauthier's argument is that agents can predict to a reasonable degree the dispositions of others, it follows that agents would also be able to identify those that lean towards a disposition to reserved maximization, and hence, they will move to protect themselves in jackpot situations.

Problem 4

That Gauthier succeeds in arguing that CM is a rational disposition or that he was successful in showing that the rationality of adopting a CM disposition carries over to the action itself does not mean that he succeeds in arguing that such a disposition can be formed. Smith frames the problem as follows:

> Even if you know that I choose CM, all you can infer from this is that I will form the intention to build and carry it out *if and only if* I predict you

> will build. You cannot infer that I will build, simpliciter. But if you cannot infer this, then you will not build. And I will not form the intention to build and then carry it out unless I believe you will build. Neither of us, knowing the other has chosen CM, has sufficient information to predict on that basis what the other will do, so neither of us can decide which intention to form and act to carry out.[184]

Response

Since the disposition to CM is conditional upon others having the same disposition, Gauthier's argument that agents would form such a disposition exposes a practical or logical contradiction. I have to form a CM disposition on the basis of your having a similar disposition and you have to form a CM disposition on the basis of my having a similar disposition. If everyone is waiting for everyone else to form the disposition to cooperate, no one will ever form such a disposition. There are at least three ways to address this problem.

Firstly, the argument that agents can form simultaneously the dispositions to do *A* and to do *B*. Stated differently, X forms a CM disposition and an SM disposition and X acts on either, depending on what disposition Y has, and Y forms a CM disposition and an SM disposition, and Y acts on either, depending on what disposition X has. This seems implausible, for it suggests that agents can simultaneously commit themselves at time t_1 to doing *A* at time t_2 and not doing *A*, but doing *B* at time t_2. Dispositions may not be ephemeral, but they are not invariant.

[184] Smith, "Deriving Morality from Rationality," *Contractarianism and Rational Choice,* p.240.

Even if dispositions are not invariant it is not quite clear how the changing nature of dispositions prevents an individual from simultaneously forming two dispositions that point in different directions and executing either at the appropriate occasion. What I have in mind about simultaneously forming two dispositions can be illustrated as follows. Suppose there is a 0.5 probability that it will rain tomorrow; I can simultaneously form, today, two dispositions that point to different directions and be able to execute either tomorrow, depending on weather conditions. I could say, "If it rains tomorrow I will stay at home" but "If it doesn't rain tomorrow I will go to school." Whatever disposition I end up executing tomorrow will depend on whether it rains or not.

As appealing as this argument is, it is not convincing. If it rains tomorrow, I will stay at home, but if it does not rain tomorrow, I will go to school. The dispositions I formed here, we might say, are 'one-direction conditional,' in the sense that I form the dispositions today, conditional on my expectation of what the weather will be tomorrow, which neither depends on the dispositions I have formed nor on which disposition I execute. But in the case of simultaneously forming a CM disposition and an SM disposition, the dispositions I form here are not 'one-direction conditional,' but, we might say, a 'two-direction conditional,' in the sense that they are conditional on your or the other person forming similar dispositions, which depends on which dispositions I have formed. Since I need you to form the dispositions before I form mine and since you need me to form the dispositions before you form yours, we would never be able to form any dispositions. Simply put, since the dispositions, say, CM and SM that X forms are 'two-direction

conditional,' formed on the basis of what dispositions Y has formed, whose dispositions is formed on the basis of the dispositions Y has formed, neither X nor Y will be able to form CM and SM dispositions.

The second argument is the argument that agents will be able to form layers of dispositions or meta-dispositions. According to the argument for layers of dispositions, agents do not adopt a disposition to CM but rather they form a disposition to be disposed to a CM disposition, just in case other agents form a disposition to be disposed to a CM disposition. As is with the argument of simultaneously forming dispositions, the argument for layers of dispositions is problematic. The appeal to second-order dispositions pushes the problem one-step back. Perhaps, the argument can be made that not everyone forms second-order dispositions. I form first order dispositions, say, a CM disposition while you form second-order dispositions, say, a disposition to be disposed to a CM disposition. But it is not clear how this addresses the worry that since I need you to form a particular disposition (whether first or second-order) before I form mine and since you require me to form the disposition before you form yours, we would never be able to form any disposition.

Thirdly, the argument that when an agent forms a particular disposition what happens is that such a disposition positively induces others to form a similar disposition. That is to say, dispositions are mutually reinforcing, such that when I form a particular disposition, I *encourage* you to form a similar disposition. This seems to be partly what the argument for layers of dispositions is getting at. The argument that dispositions are mutually reinforcing does not take the disposition a

person forms to be dependent on the disposition the other person forms. Rather, it takes the disposition a person forms to induce a similar disposition from the other person. On this view, the person who forms the disposition to CM does not form the disposition on the basis of being able to identify that the other person is willing to form a similar disposition. The person forms the disposition on the basis of her knowledge of what the person she is interacting with *is like*.

The argument that forming a disposition induces others to have a similar disposition is promising. Given that the whole point about forming dispositions is to make cooperation possible and to guarantee higher benefits for everyone, it stands to reason that an agent would be encouraged to follow others in forming dispositions congenial to cooperation. But what does it mean to say that a person forms a disposition on the basis of her knowledge of what the person she is interacting with *is like* and that this disposition induces or encourages the person to form a similar disposition?

To answer this question we have to exploit the argument for the transitive relation between a principle and a disposition, between a disposition and an action, and between a principle and an action. The argument that a person chooses a disposition because it expresses a particular principle and having the disposition is expressed by the action that person chooses provides us a veritable platform to make the case that dispositions are mutually reinforcing. To illustrate, suppose your reason for walking your dog everyday relates to your position on equality. Let us say you have this as your principle, the "principle of the equality of human and non-human animals," that is, you believe that non-human animals should be treated

as kindly as humans are treated. Suppose also, that both you and I are contemplating entering into some contract, let us say to take care of each other's pet when the other is away on vacation. Under these circumstances, if we accept that P = D, D = A, therefore P = A, then each of us would expect that the other person will not only be congenial to treating animals kindly, but would always seek to do so as best as possible.

My awareness that you have the "principle of the equality of human and non-human animals" provides me additional reason or shall we say, a simple guarantee, so to speak—besides of course, the reason that both of us benefit from cooperation—to form the disposition to take care of your pet when you are away. The simple guarantee is etched in and expressed in your prior actions, namely, in your relationship with and treatment of non-human animals. My forming the disposition to look after your pet during your vacation induces and encourages you to form a similar disposition to take care of my pet when I go away on vacation.

As promising as this argument is, it suffers from one problem. It assumes that principles and dispositions have some quasi-magical property. It seems right to claim that one can form a particular disposition if one knows what principles others act upon, assuming that the disposition expresses a particular principle. But it doesn't follow from this that one's adoption of a particular disposition would necessarily induce the other to form a similar disposition. People do change their principles, and with it their dispositions, and sometimes when they do, there may not be the occasion for others to observe the actions that express these principles or dispositions, and hence may not be encouraged to form similar dispositions.

Suppose, after reading some book on the 'exotic habits of non-human animals,' you now believe at this moment that it is not good to treat equally humans and non-human animals, or that it is good to only treat kindly some non-human animals. Your principle, we might say, has changed here and now, but I have no way of knowing this *at this time*. By coincidence, my pet does not belong to the group of non-human animals that you believe ought to be treated kindly. You are still kind to your dog, but this would not indicate that you are operating on the "principle of the equality of human and non-human animals." Your pet, and not mine, belongs to the group of non-human animals that is to be treated kindly. In this circumstance, even though I form the disposition to care for your pet when you are away, you may not be induced and encouraged to form a similar disposition to care for my pet when I go away on vacation.

Problem 5

Problem 5 challenges some of the assumptions Gauthier employs in Mb(CM)A. The criticism goes this way: Gauthier appeals to assumptions that are *ad hoc* and unrealistic; therefore, he undermines his project of grounding morality on rationality. There are two parts to this criticism. The first part is raised by DeBruin:

> [I]t is a "*contingent* matter whether agents are sufficiently translucent for Gauthier's purposes. It is not an essential feature of rational agents that they are incapable of deceiving others about their dispositions. Insofar as

> translucency is a contingent feature of agents, it is possible that some agents will be opaque."[185]

In other words, since Gauthier's methodology requires that he appeal to an essential feature of human beings to commit all humans to morality, he cannot appeal to any contingent feature. Translucency of disposition is a contingent feature and a view that claims that agents are able to predict sufficiently the character of others will not apply to "*any* sufficiently opaque agent." Therefore, Gauthier has not provided enough justification as to why one must adhere to morality.

The second part of the criticism is that even if it is the case that people have translucency as an essential feature, agents may "strive to appear to be translucent while, in fact, they are opaque" so as to reap the fruit or advantages of living within a moral community. As Sayre-McCord puts it, "they will develop winning smiles, travel with a glowing reputation, and cultivate an honest manner"[186] all in a bid to deceive others and make them misjudge their true character. Like DeBruin, Sayre-McCord therefore, concludes that Gauthier fails to establish that rationality always requires agents to dispose themselves to comply with the requirement of morality. He writes:

> [S]imply being a translucent person is not enough to justify the choice of a moral character. If others in one's community are relatively opaque, or if a sufficient number are unvarnished egoists, then being moral would simply set one up as a sitting duck. The choice of a moral character is

[185] Debra A. DeBruin, "Can One Justify Morality to Fooles?" in *Canadian Journal of Philosophy*, vol. 25, no 1, March 1995, p.20.
[186] Geoffrey Sayre-McCord, "Deception and Reasons to be Moral," in *Contractarianism and Rational Choice*, p.191.

> rational, then, only if one has reason to think one is a (sufficiently) translucent member of a community of (sufficiently) translucent moral people.[187]

Response

Gauthier's strategy might be to deny that translucency of disposition is a contingent feature. He might claim that translucency of disposition is the default from which deviations such as opacity and semi-opacity of disposition are explained. There are two ways of interpreting this claim.

First, as descriptive, i.e. that our characters or dispositions are translucent—they are neither 0 nor 1, but lie between 0 and 1. Given that our dispositions are naturally translucent we move away from the default when we attempt to conceal them or to appear opaque. Those who attempt to cloak their dispositions for the purpose of deceiving others are engaging in wish fulfillment, i.e. they are hoping and thinking that if they just try hard enough to conceal and cloak their true character they would, perhaps at this time, reap the benefits of doing so.

Second, the claim can be interpreted as normative, i.e. that we are rational to the extent we expose our character or to the extent our dispositions are sufficiently accessible to others. A rational agent may not choose to conceal his or her character since to do so is to display a measure of irrationality. On this viewpoint, any deviation from the requirement that one's character be accessible to others is clearly irrational and since irrational people are not the sorts of beings that Mb(CM)A is concerned with, those who cloak their disposition are excluded from

[187] Ibid, p.191.

the sphere and horizon of rational morality. Mb(CM)A can be interpreted as appealing to both the descriptive and normative views.

Suppose we agree that translucency is an essential feature, how might it be specified? I suppose there are a number of ways of doing this. One way might be in terms of the transitivity relation argument: P = D, D = A, therefore P = A. We might take the transitivity relation to be a cardinal feature of rationality, the same way coherent preference is a cardinal feature of rationality. On this view, P = D, D = A, therefore P = A describes how people are. That is, our actions follow from and express the dispositions we have, which may express the principles we have. And since a disposition is disclosed by a principle and expressed by an action, the disposition of a rational person is invariably translucent. On this view then, translucency of disposition, like preference coherency, is the default of the ideal rational agent that explains every divergence such as opacity or semi-opacity of disposition or irrationality.

Preference coherency forbids me from preferring *z* to *x* if I do indeed prefer *x* to *y* and *y* to *x*. If I prefer *x* to *y* and *y* to *z*, then I must choose as my preference suggests, i.e. choose *x* over *z*. The transitivity relation of preference coherence could be understood as either a descriptive or a normative claim of rational preference. As a descriptive claim, it describes how people's preferences are. And as a normative claim, it stipulates that an agent must not hold incoherent preferences. Nevertheless, people may deviate from both the descriptive and normative claims and choose to hold incoherent preferences. When they do this, because of the conflicts of preferences, they reduce their utility profile and

compromise their life-plan. In the same way, people may choose to conceal their dispositions. And since they would not be recognized, they would be assumed to be cheaters, and hence, would be excluded from cooperative schemes that are mutually advantageous and that provide greater benefits.

The thought that opacity comes at a great cost seems self-explanatory. In any scheme of cooperation that is mutually beneficial and where knowledge of what people will do is crucial, those whose characters are hardly recognizable or who are unidentifiable because they have poker faces are generally distrusted, and ensuingly excluded from such scheme. Therefore, individuals may not want to be mistaken as SMers or unidentifiable or misrecognized because they choose to mask their identity. There is the other side of the coin: because the dispositions of agents are sufficiently revealed to other agents, those not disposed to cooperation would be avoided in situations where interactions are mutually advantageous. Since there is a greater chance of being correctly identified, agents will be motivated to reveal their identify so as to benefit from the opportunities not open to those who choose not to reveal their identity.

But this type of argument, according to Sayre-McCord, assumes too much. It "assumes implausibly that we live in a community that will raise significantly the risk of deception, and it plays on considerations of what would happen in a community of fully rational agents (a community we surely don't live in now)."[188] And in view of the fact that "people are, in fact, both ignorant and irrational" any argument that "assume otherwise is not of practical interest."[189] Similar

[188] Ibid, pp.193, 194.
[189] Ibid, pp.193, 194.

considerations underlie Vallentyne's criticism of Gauthier's assumption of mutual unconcern. He says, "Rationality requires that one use realistic assumptions…. By making the legitimacy of norms depend on what we would agree to if we had preferences that we do not in fact have, Gauthier undermines the rationality of the agreed-upon norms."[190] Or as Braybrooke puts it, "Without these assumptions, Gauthier would not have much chance of reaching morally convincing results… From the aims of rational agents under other assumptions—assumptions that left predatory behavior unchecked and allowed intimidation to overshadow voluntary entry into the bargain—a convincing morality very likely would not be deduced…"[191]

If we accept this objection, then Gauthier's overall strategy of idealizing his agents is undercut and subsequently any morality built upon the assumption of rational actors would be suspect. Obviously, it is one thing to idealize one's agents as Gauthier does and it is another thing to map them to real agents in our societies. It is not just that there may be no real fit between Gauthier's agents, it is that there can *never* be such a fit because Gauthier assumes capacities that only "super rational" people possess. If there is no such fit, then Mb(CM)A, according to Braybrooke, "cannot be used for real world policy-making." He writes:

> [T]he degree of technical perfection to which Gauthier has brought social contract theory… has been gained at the expense of depriving the theory of any possibility of effective application. Its demands for information are fantastic—too fantastic ever to be met or even to allow the theory to be

[190] Vallentyne, "Contractarianism and the Assumption of Mutual Unconcern," in *Contractarianism and Rational Choice*, pp. 72, 74.
[191] Braybrooke, pp.755, 756.

> used as a guide to improvements within the reach of present social policy.[192]

Note that the objection challenges the descriptive viewpoint: agents do not have the capacities (preferences, information) ascribed to them in Mb(CM)A. At the extreme, Gauthier could respond to it by following Hume who argues that if rational or virtuous individuals find themselves outstripped by a community of irrational people—monsters and ruffians then the rational thing to do is to suspend any particular regard for justice and morality, and simply "act like the Romans do in Rome." For if CM, which "disposes us to justice, will indeed be of no use to us, and we must then consult only the direct dictates of our own utilities." In a world of Fooles or a world that does not satisfy the circumstances of justice it would not pay to be a CMer or to comply with one's agreements. In such circumstances, it would not be rational to be moral. In a world of monsters and ruffians or in a society dominated by Fooles it will pay an individual to be a Foole.[193] If indeed, it turns out that we live in a community of ruffians and monsters or among people who are not disposed to morality, then morals cannot be by agreement. The point is that moral constraints are valuable and effective only in a community where a good number of the people are disposed to morality. That morality of any identifiable sort is not to be expected in a community of ruffians and predators is well attested to by the Hobbesian state of nature of war of all against all.

Gauthier does not have to embrace this extreme view. There are two ways he might respond to the objection. Firstly, he could say that the objection underlies

---

[192] Braybrooke, "Social Contract Theory's Fanciest Flight," *Ethics:* An International Journal of Social, Political, and Legal Philosophy, vol. 97, no 4, July 1987, p. 751.
[193] MbA, pp.181-182.

a misleading caricature and view of the human condition, a view that mistakenly assumes that humans do not possess many of the capacities he ascribed to them. This response claims that it is not the case that humans do not have these capacities. Descriptively, humans have these capacities and if they do not recognize that they possess them they can be persuaded by philosophical reflection to understand that they do in fact possess them. On a side note, it is true that we do not live in a society of saints, but equally true is that we do not live in a society dominated by Fooles—people seem to be for the most part rationally and morally disposed.

Secondly, Gauthier might agree that there are indeed these empirical facts—the fact that different people do manifest all sorts of weaknesses, imperfections, and irrationalities—that his theory ignores, but then deny that this renders his account implausible. Again, that people have incoherent preferences is not an indication that rational people do not hold coherent preferences or that rationality does not require that they hold coherence preferences. It is true that people may hold incoherent preferences due to weakness, imperfection, ignorance or any other reason, but this does not make the preferences themselves rational. Above and beyond this, one thing that morality as an enterprise does or that we hope it should do is to challenge behaviors that subvert the moral community. What is morality if not a set of constraints on behavior, and why limit behavior if not for deviations of irrationality like weakness and ignorance. Gauthier states this about the constraining nature of morality:

> This rationale for agreed constraint makes no reference to the content of anyone's preferences. The argument depends simply on the *structure* of interaction, on the way in which each person's endeavor to fulfill her own preferences affects the fulfillment of everyone else. *Thus each person's reason to accept a mutually constraining practice is independent of her particular desires, aims and interest, although not, of course, of the fact that she has such concerns*…. Morality is not to be understood as a constraint arising from reason alone on the fulfillment of nonrational preferences. Rather, a rational agent is one who acts to achieve the maximal fulfillment of her preferences, and *morality is a constraint on the manner in which she acts, arising from the effects of interaction with other agents*.[194]

The point is clear: Gauthier assumes that we are rational enough to accept the constraints that morality imposes. The initial bargaining position, he argues, is constrained by the "Lockean Proviso" (I discuss this in the next section). The Lockean Proviso has several implications, foremost of which is that humans have certain rights—rights as property holders—that must be respected if the market and any scheme of cooperation are to be fully operational. Since those that have the tendency to exploit others for gains are only able to beneficially and effectively exploit for gains in a state of rational constraints—a point that the Hobbesian state of nature has brought to our attention—there won't be predators, ruffians and monsters. For anyone to effectively pursue his or her interests, whatever these are, he or she must respect the rights of others as property holders. These rights, which prohibit victimization by others or the violation of the rights of others are not part

[194] Gauthier, "Why Contractarianism," pp.23, 24 (my emphases).

of the contract but are necessary for it to get off the ground. How did Gauthier argue for these presocial, pre-contractual rights? I now turn to this.

3.1.3 The Contract Problem: The Natural Baseline and the Proviso

In the penultimate section (3.2.1), I examined Gauthier's argument for the MRC principle as a solution to the bargaining problem, which is the problem of selecting among a number of possible but mutually incompatible distributive schemes of the cooperative surplus. In my discussion of the distributive principle I deliberately left out one important issue: the issue of how initial factor endowments, i.e. rights as property holders, arise. Put differently, I did not consider the exercise of natural endowments and talents and the use of this in acquiring resources or the question of how people come to have rights as property holders that are essential to the bargaining process. I now want to examine this issue in this section in connection with the third sub-theory of Mb(CM)A, i.e. the contract problem.

We can categorize the major contending theories of distributive justice by the way they specify property acquisitions or rights. As we saw in chapter two, Rawls takes property rights as part of the "social product" and emphasizes their irrelevance in determining the cooperative surplus. Property rights are not natural and are to be regulated by the difference principle. Libertarians like Nozick take the contrary position by defending pre-contractual property rights and arguing for their significance in determining desert. Although Gauthier, like Rawls, develops a rigorous and systematic theory of distributive justice by drawing on the resources of rational choice theory, he rejects the strong egalitarianism and drastic

redistributionism demanded by Rawls. Gauthier sides with Nozick by defending presocial property rights, arguing for their significance in the market and in the scheme of cooperation. However, he differs from Nozick on the moral status of pre-contractual natural rights.

For Nozick, pre-contractual rights are natural because they have an independent moral appeal for moral agents. That is, they are side-constraints and form the basis of assessing what right holders can do and what others, including the state can do to them. However, for Gauthier, pre-contractual rights are natural in the sense that they are defended merely as necessary pre-conditions for the social contract. Given the importance that property rights plays in Mb(CM)A, Gauthier has to provide an argument for their emergence. He has to explain how people come to possess the resources that form the core of the bargaining process, i.e. how people come to own bits of property in the natural world. Gauthier thinks he can provide such an account. He begins by identifying the most appropriate initial position from which agreements are produced. The initial position is defined by specifying what types of property rights that individuals possess and the sense in which these are taken to be legitimate.

Gauthier follows Locke and Nozick in identifying personal property rights and appropriations by the Lockean proviso.[195] Gauthier provides a justification for constraining appropriations by the proviso in the following long passage:

---

[195] The sensitivity condition is the Lockean proviso that prohibits acquisitions that worsen the condition of others. Locke's explicit reference to the proviso is in section 27 of Two Treaties of Government where he says, "Though the earth, and all inferior creatures, be common to all men, yet every man has a *property* in his own *person:* this no body has any right to but himself. The *labour* of his body, and the *work* of his hands, we may say, are properly his. Whatsoever then he removes out of the state that nature hath provided, and left it in, he hath mixed his *labour* with, and joined to it something that is his own, and thereby makes it his *property.* It being by him removed from the

> The initial bargaining position must be non-coercive. But must we go further in constraining natural interaction, in so far as it determines the basis of market or cooperative interaction? We shall argue that the terms of fully rational cooperation include the requirement that each individual's endowment, affording him a base utility not included in the cooperative surplus, must be considered to have been initially acquired by him without taking advantage of any other person—or, more precisely, of any other cooperator. Otherwise *those who consider themselves taken advantage of in initial acquisition will perceive society as unfair*, in demanding payments from them without offering a compensating return, *and will lack sufficient reason to accept market arrangements or to comply voluntarily with cooperative joint strategies*.[196]

Rightly, Gauthier recognizes that issues of impartiality and stability would affect the contract if an appropriate initial position is not specified. If those who come to the bargaining table have previously but "unjustly" acquired resources, it is likely that the victims of such injustice would find unfair and unacceptable the bargaining process and the agreement that they generate. And because they will find the bargaining process unfair, they will lack sufficient reason to support market arrangements or to comply voluntarily with agreements. Suppose that before Abel and Mabel invested in the money market Abel had taken, by force or fraud, some of Mabel's goods, which he then sells to someone else. Suppose also that part of the money invested in the money market by Abel includes the proceeds from the sale of the goods he took from Mabel. It is more than likely that Mabel

---

common state nature hath placed it in, it hath by this *labour* something annexed to it, that excludes the common right of other men: for this *labour* being the unquestionable property of the labourer, no man but he can have a right to what that is once joined to, at least where there is enough, and as good, left in common for others."

[196] MbA, p.200, my emphases.

would consider his initial factor endowment illegitimate and consequently refuse his full claim to the cooperative surplus.

Peter Danielson has criticized Gauthier's identification of the initial position with pre-contractual property rights. An appropriate initial position, he argues, need not incorporates property rights since these are not necessary for cooperation, agreement and bargaining. He states:

> [A]ny advantage taken in appropriation is open to rectification in the social bargain, where the morality of appropriation is ultimately settled.…To subordinate appropriation in this way is the alternative we proposed.…[we] treat property claims as subject to social agreement.…Cooperation may be agreed to – even *bargained* to – so long as agents are defined in any number of less definite ways, including appeals to equality and non-coercion. The contractors need some pre-contractual individual rights but they do not need fully developed property rights.[197]

Danielson's argument leans towards the view that contractors care less or are apathetic towards the history of appropriations and that they consider the status and history of appropriations irrelevant to how things are now, to how they wish to define their relationships with fellow contractors, and to the direction the bargaining process should take. However, to say that contractors have a slipshod and apathetic view of the history of appropriations is somewhat misleading. To

[197] Danielson, "The Invisible Hand of Morality," in *Canadian Journal of Philosophy*, pp. 368,369. While Danielson criticizes Gauthier's account of pre-contractual property claims from the left, i.e. that the gains of appropriations prior to the contract ought to be subjected to social agreement, Narveson criticizes Gauthier's account of pre-contractual property claims from the right, i.e. that the gains of appropriations prior to the contract ought to determine desert in the market and in the scheme of cooperation rather than being subjected to the moral force of the MRC principle. See Narveson, "Gauthier on Distributive Justice and the Natural Baseline," in *Contractarianism and Rational Choice*, pp. 136-146.

suggest this is to suggest that notwithstanding the antecedents of appropriations, any agreements would *elicit* the willing cooperation of *all* contractors.

Most certainly, from the point of view of individual interactions, there might be some—perhaps, those with little appropriation—who favor a scheme that excludes appropriation or what contractors bring to the bargaining table from the bargaining process. But to define the *entire* bargaining and contract process and the contract from the perspective of those with little appropriation is to define, like Rawls does in JaF, the contract from the standpoint of a particular group of people, namely, the least advantaged or worst-off group. To define the contract from the standpoint of the worst-off group is to ignore those who have *legitimate* property claims. As Gauthier rightly notes, those who have justly acquired their property will feel "taken advantage of if initial appropriation is not permitted to influence, however robustly or slight, the social bargain."[198]

In the absence of fully defined property rights and in preventing the effects of appropriation from determining interactions, and what people get in society, agreements can hardly be considered impartial. To ignore what may have happened prior to cooperation, agreement and bargaining is to ignore something fundamental about contractors, their natural capacities, and the effects of these on appropriations. If the money that Abel invested in the money market was not taken by force or fraud but rather was legitimately acquired, i.e. constrained by the Lockean proviso, he would be right to insist that he is entitled to the full share of the cooperative surplus that part of the money produced. And any move by Mabel

[198] Gauthier, "Moral Artifice: A Reply by Gauthier," p.412.

to deny him this would have no rational or moral justification, and would be considered by him unfair.

A similar argument can be made were we to consider the issue of appropriation from the Archimedean standpoint. A choice made from the ideal standpoint would consider appropriations legitimate in the same way appropriations are considered legitimate from the standpoint of individual interactions. Danielson mistakes the Archimedean chooser for the rational chooser in JaF, who chooses behind the veil of ignorance. The ideal actor chooses impartial and beneficial interactions and in choosing rationally the actor, Gauthier argues, "chooses the proviso."[199] The rational chooser in JaF identifies with *no one* in particular. She has no individuality and so chooses to regulate appropriations by the difference principle. But the Archimedean chooser chooses in full view of everyone's individuality. She chooses as if she *were* every person. In choosing as every person, she is constrained by everyone's individuality and must necessarily recognize everyone's capacities, talents, attitudes, preferences and what they produce. She reproduces her knowledge of the personal characteristics and circumstances of everyone in her choice and so chooses the proviso.

In specifying the initial position, Gauthier follows Nozick in arguing that the Lockean version of the proviso is too strong. For Nozick, the crucial point associated with Locke's proviso "is whether appropriation of an unowned object worsens the situation of others."[200] Locke's proviso with its emphasis on appropriation that leaves enough and as good for others, according to Nozick, is

[199] Gauthier, see MbA, pp.259, 260.
[200] Nozick, *Anarchy State and Utopia*, p. 175.

meant to ensure that any appropriation does not worsen the situations of others. Making someone worse by appropriation, Nozick says, can be done in two ways: the stronger sense and the weaker sense. Thus, we need to distinguish the sense that Locke has in mind. The two ways that someone can be made worse by appropriation are:

(1) when a particular appropriation makes an individual lose an opportunity to improve her situation, for example, your appropriation of the best arable land in a desert island deprives me of the opportunity to do same.
(2) when a particular appropriation prevents an individual from being able to any longer freely use what she previously could, for example your appropriation of a communal fish pond or park, prevents me from freely using either of them again.

According to Nozick, while the stronger sense of the proviso prohibits both, the weaker sense prohibits only the second.[201] In modifying Lockean proviso, Nozick takes the weaker sense, i.e. all appropriations except those that violate the second. The weaker version of Locke's proviso specifies the conditions or situations wherein appropriation is both allowed and prohibited. Although appropriations might violate the proviso in general (including those that fall in the second category) by leaving people worse off, i.e. in a situation worse than the baseline, an appropriation might still be allowed, as long as compensation, according to Nozick, is paid to those who have been made worse by the appropriation.

[201]Ibid, p.176.

The weaker sense of Locke's proviso (which we will refer to as Nozick's Proviso, or simply as NP) prohibits worsening the condition of others, whether by predation, plundering, depredation or in other ways. According to Nozick, such prohibitions—and indeed any account of a proviso that incorporates the weaker sense of the proviso—are important for any satisfactory theory of distributive justice. As he writes, any "adequate theory of justice in acquisition [as well as a theory of justice in transfer] will contain a proviso."[202]

In spite of Nozick's weakening of Locke's proviso—prohibiting cases of appropriations that deprive an individual free use of an object that she previously did use—Gauthier thinks it (NP) is still too strong. It is too strong because it requires that where the choice of plundering of a property is between yours or mine, I can allow mine rather than yours to be plundered.[203] NP, Gauthier claims, allows me to worsen my situation at the expense of another in cases where either yours or my position is to be worsened.

Weakening NP then, according to Gauthier, requires that we specify worsening in terms of interaction, such that it forbids worsening the situation of another except to avoid worsening one's own. Recall, that Gauthier's agents are rational utility maximizers and nontuistic, that is they are mutually unconcerned. Since he assumes this of his agents—whether those in natural interactions or those engaged in schemes of cooperation—invoking a weaker sense of NP is meant to provide them the space whereby they can maximize their benefits or utility within

[202] Ibid, p.178, 179.
[203] Otherwise put, in situations of dire scarcity or where someone's or everyone's property is to be appropriated.

limits that are considered fair and just.[204] Gauthier's weakened version of the proviso "prohibits bettering one's situation through interaction that worsens the situation of another."[205] This view of the proviso, he claims, represents better Locke's intuition, where one's preservation takes justifiable precedence over that of others.

Locke's intuition regarding appropriation is circumscribed by his views of two obligations regarding preservation, which seem to warrant the kind of revision of the proviso that Gauthier is proposing. For Locke, humans have two basic obligations: an obligation to God under the natural law of self-preservation and an obligation to preserve others or humankind in general, with the former trumping the latter. The realization of the preservation of humankind in general is dependant on how consistent it can be made to fit with the natural law of self-preservation. Given, then, Locke's view of obligations, it seems right to conclude that in situations where the choice of survival or plundering is between one's own and others, one's self preservation takes justifiable precedence over that of others.

Gauthier's motivation in specifying the initial position defined by the proviso is not significantly different from Nozick's. For Nozick, the proviso specifies whether one's holdings and the transfers that are made are just and legitimate. This consideration is also important for Gauthier, who believes that the proviso justifies pre-contractual personal rights and acquisitions and ensures that the entire contract is not tainted by unfairness. He writes:

---

[204] Gauthier says this of the two assumptions of the proviso, "The proviso is intended to apply to interaction under the assumptions of individual utility-maximizing rationality and mutual unconcern. Each person is supposed to chose a strategy that maximizes his expected utility, unless specifically forbidden by the proviso to do so," MbA, pp. 205-206.
[205] Ibid, p.205.

> But we noted explicitly that fair procedures yield an impartial outcome only from an impartial initial position. And it is equally true that rational procedures yield a rationally acceptable outcome only from a rationally acceptable initial position.[206]

However, unlike Nozick, the proviso plays a bigger role in Gauthier's theory of distributive justice. It is a necessary precondition for the possibility of the contract, market interactions, and any scheme of cooperation. The proviso for Gauthier moralizes and rationalizes the state of nature insofar as this is construed as leading to civil society. In this way, the proviso, for Gauthier, is forward looking and not backward looking like it is for Nozick. For if there were no possibility of leaving the state of nature, the proviso would be otiose since it would be irrational for anyone to constrain their behavior without others doing same. So, if there were no hope of civil society emerging from the state of nature there would be no need to adhere to the proviso, since no one benefits by independently and unilaterally constraining their behavior.[207]

Accepting the proviso as a constraint on interaction, then, is dependent upon the possibility of society emerging. What is the probability that agreement would arise or that the market or civil society would emerge? Gauthier thinks it is quite high given the fact that the state of nature is a failed state, a state of suboptimality. Since the state of nature is not optimal nor Pareto-efficient it would be irrational for humans to remain in the state of nature where they accept no constraint on their interactions. He notes:

---

[206] Ibid, p.191.
[207] Ibid, p.193.

> Without the prospect of agreement and society, there would be no morality, and the proviso would have no rationale. Fortunately, the prospect of society is realized for us; our concern is then to understand the rationale of the morality that sustains it.[208]

Circumscribing appropriation vis-à-vis the initial condition by the proviso therefore, prevents the outcome of the bargain from being tainted by the effects of coercion, fraud, predation, parasitism, and freeriding. It makes it rational for agents to keep the terms of the bargain and ensures that issues of stability or compliance do not loom large or weigh heavily on the contract.

Now, suppose on a small island there exists a community of masters and slaves.[209] The slaves serve the masters and perform various tasks grudgingly and carelessly under the threat of whips and chains. Due to this coercive apparatus, both the masters and slaves find themselves in a situation of suboptimality. The masters have fewer resources for real pleasure because they spend so much in providing and maintaining the coercive apparatus needed to keep the slaves in control. In addition, because the slaves work grudgingly and carelessly their productivity is not optimal. The slaves on the other hand are forced to work and they receive little wages and living allowance because the masters do not have enough resources to pay them higher wages.

What would be the outcome if we suppose that the slaves decide to work as willing servants, that is, if the coercive apparatus that keep the slaves in check are dismantled? Without any coercion, we would suppose that the slaves would have

---

[208] Ibid, p.193.
[209] Gauthier uses this example to set up his discussion of the proviso in chapter VII; see MbA, pp.190-199.

more wages and, hence, would have a better or higher standard of life. The masters would also have more resources and would also have a better or higher standard of life because they would not need to maintain the coercive apparatus necessary to keep the slaves in check.

Let us call the first situation (the slave situation) state of affairs A and the second situation (willing servant situation) state of affairs B. In the two situations of the master and slave tale, Gauthier thinks that our intuition would seem to suggest that B is better than A. But unfortunately even though B seems better, it would be hardly achievable because the "bargain [moving from A to B] was coercively based."[210] And as the prime minister (one of the ex-slaves) of the new administration (state of affairs B) says after being sworn in, they (former slaves) are not about to become willing servants (move voluntarily from state of affairs A to state of affairs B). They are not prepared to voluntarily comply because the fruits of cooperation—even if they respect the MRC principle—are a bargain that is based on a coercive initial position. Gauthier thus concludes from the account that "implicit in the prime minister's remarks in our cautionary tale is the claim that it is rational to comply with a bargain, and so rational to act co-operatively, *only if its initial position is noncoercive"* or unfair.[211]

Since Gauthier's version of the proviso (let us call it GP) "prohibits bettering one's situation through interaction that worsens the situation of another," the base point for judging or determining how the actions of individuals affect others (bettering or worsening) is in terms of (their absence or presence).

[210] Ibid, p.191.
[211] Ibid, p.192, my emphasis.

Therefore, what determines my bettering or worsening your situation is by comparing what I actually do to you (the state of affairs I bring about while interacting with you) with the state of affairs that would have occurred in my absence. In parsing out how GP incorporates the two assumptions of individual utility maximizers and mutual unconcern, Gauthier specifies three possible sets of strategies open to individuals, which are ranked in the scale of 1 to 3:[212]

> S1: Strategies that give an individual an expected utility (EU) that is greater than or equal to but not lesser than an EU in the absence of interaction.
> S2: Strategies that give an individual an EU that is greater than or equal to but not lesser than an EU in the absence of interaction, but gives someone else an EU less than an EU in the absence of interaction.
> S3: Strategies that give an individual an EU less than an EU in the absence of interaction.

According to Gauthier, an individual faced with the above three sets of strategies would choose in this order: S1, S2, S3. She chooses S2 only if S1 is empty, and she chooses S3 only if S2 is empty.[213]

What about if the proviso is violated? Then Gauthier, like Nozick, holds that compensation ought to be paid.[214] What determines whether the compensation is full compensation or market compensation is whether the cost that X imposes on Y is a displaced one or if X prevents Y the use of goods that Y previously had use of. If the former, the compensation has to be full but if the latter, then what is

---

[212] Ibid, p.206.
[213] Ibid, p.206.
[214] Ibid**,** p.211; Nozick says this as well in *Anarchy State and Utopia* about compensating others when appropriations violate the proviso, "Someone whose appropriation otherwise would violate the proviso still may appropriate provided he compensates the others so that their situation is not thereby worsened; unless he does compensation these others, his appropriation will violate the proviso of the principle of justice in acquisition and will be an illegitimate one," p. 178.

required is market compensation. While full compensation leaves Y without any net loss in utility, market compensation gives her the opportunity to share in the benefits that accrue to Y, who by seizing the goods, exclusively appropriates them.

Displaced cost has to be compensated because the cost X imposes on Y is necessary to the benefit she receives from interacting with Y. But if the worsening of Y by imposing some cost on her is incidental and unnecessary to the benefit X receives, then X has not violated the proviso and does not need to pay any compensation. The reason that X does not need to pay any compensation to Y in cases of incidental costs is that X does not better her situation through interaction with Y even though the condition of Y is worsened by the cost that X offloads on her.

3.1.3.1 The Proviso and Different Senses of 'Bettering and Worsening'

To give us a better understanding of NP and GP as a foundation for pre-contractual property rights, let us finely distinguish the different senses or strategies of bettering and worsening provided by the proviso in general.

S1: X betters her condition by bettering Y's condition.
S2: X betters her condition by leaving Y or others neutral.[215]
S3: X betters her condition by worsening Y's condition.
S4: X worsens her condition by bettering Y's condition.
S5: X worsens her condition by leaving Y or others neutral.
S6: X worsens her condition by worsening Y's condition.

[215] Neutral in this context means neither worsening nor bettering someone's condition, that is, my interaction with you leaves you in the same condition, neither better nor worse.

Both NP and GP rule out S1 because requiring X to better the condition of Y or others while bettering hers would be to require that she allow others to be parasites on her or to require her to breed freeriders.[216] Of course, X can choose to be *nice* to Y, i.e. X can choose to improve Y's condition. However, the proviso does not require her to do so. It only requires her not to worsen Y's condition. Since the proviso wants to prohibit cases where one betters one's condition through interactions that worsen the conditions of others, adopting S2 as a strategy would seem not to violate both NP and GP. If my interaction with you leaves you no better-off (neither worse nor better), although I am better-off, I would have offloaded no displaced cost onto you. The obvious case of proviso violation would be S3. S3 violates NP and GP in the sense that X offloads a displaced cost onto Y; in short, X freerides on Y's back. Unnecessarily worsening of the other party's condition at the expense of benefiting oneself or bettering one' s condition requires that one pay some compensation to the worsened party.

S4 and S5 are ruled out by GP but not by NP. By tweaking NP to accommodate the priority of individual interests and preservation over the interests of others, GP is able to take care of acts that worsen one's condition by leaving others better-off or neutral. Since the assumptions about individuals is that they are rational utility maximizers and mutually unconcerned, they would *ipso facto* choose strategies that maximize their utility profile as long as the proviso does not prohibit those strategies.

[216] In MbA Gauthier says this of freeriding, "To require that, as a condition of bettering one's own situation, one must better that of others, would be to require that one give freeriders," p. 206.

Remember that underlying Gauthier's modification of NP and the three possible sets of strategies is the idea that the individual chooses to worsen your condition rather than hers where the choice is that of worsening her condition or yours. By taking care of S4 and S5, GP and not NP, seems to explain better situations that involve conflict of interests—where the interests of the individual conflicts with those of others—and scarcity—where there exist very few resources for survival. For if interaction among individuals presupposes the assumptions of individual utility-maximization and nontuistic interests, then in cases of conflicts of interest and scarcity, individuals as utility maximizers would be disposed to rank their interests higher than those of others. GP, and not NP, thus seems to fit better with the assumptions underlying not just the strategies of S4 and S5, but also the proviso in general.

However, both NP and GP do not accommodate S6. Since both cash out expected utility in terms of worsening someone in comparison with bettering others, they do not show whether the proviso is violated by individuals who in worsening their condition make worse the condition of others. To use Gauthier's example,[217] suppose both of us live as fisherfolk along the bank of a river. I live upstream from you and occasionally use the river for the disposal of my wastes. Even though I kill many of the fish downstream (your part of the river), I do not violate the proviso, according to Gauthier, for although I worsen your situation in

[217] Ibid, pp.211, 212. In the example of the fisher folk that Gauthier uses he is more concerned with showing the various ways that interaction can take place between two individuals (of course the example can be extended to cover situations that involve more than two individuals) and how their conditions can be worsened in relation to each other. I use the example here to show how we can also incorporate situations where two individuals who interact both lose out for the benefit of someone else, i.e. a third party. My argument here is that both NP and GP do not take care of cases like this.

relation to what you would expect in my absence, I do not benefit at your expense. I do not benefit at your expense because killing your fish (due to the disposal of my wastes) does not better my situation through interaction with you. I have not offloaded a displaced cost on you since the cost I impose on you is incidental and not necessary to the benefits I receive. But if I were to do fish business with you, since your exchange power would depend on the fish you have, the killing of your fish through the wastes I disposed in the river would have through interaction worsened your situation. In this case, I have violated the proviso, because I bettered or improved my situation through interaction with you.

Let us complicate the story a bit to bring out the basic point about S6. Suppose in the account above we decide not to do fish business with each other because the fish in our river are of the same kind and species, but rather we choose to do fish business with another fisherfolk in another river, let us call her Mabel. By disposing my wastes upstream, it kills the best of my fish as well as yours, and so weakens our exchange and bargaining power with Mabel. No doubt if we choose to trade with Mabel, I would have worsened your situation as well as mine. It seems that in this situation even though I have worsened your situation I have not violated the proviso, because I have not bettered my situation through interaction with you.

But suppose that Mabel is my friend and because of having better fish and better exchange with us she becomes a very prosperous fisherfolk. Although it may be said that I have not benefited (materially or financially) from the fish business, it may not be said that I have not benefitted psychologically, especially if Mabel's

wealthy and prosperous state is a cause of delight for me.[218] What this shows is that because the proviso does not prohibit S6, individuals can worsen the situation of others as well as theirs in order to better not their situation but those of others (friends, neighbors, mates, colleagues, family members, etc.). It seems that the problem with the proviso is that in defining legitimate and just interactions (appropriation and transfers) in terms of bettering and worsening, it defines them too narrowly.

Besides the problem that S6 creates for the proviso, Hubin and Lambeth have argued that the proviso allows us to worsen the position of others to an arbitrarily large degree to avoid worsening our own only slightly.[219] Suppose that as one of only two slave owners on a desert island, I am horribly cruel and abusive but not as bad as the other slave owner. Suppose further that were I not around the other would own all of my slaves. Now, although I better my own position through interaction with my slaves, it cannot be said, according to GP, that I worsen their situation. For they are better off with me around than they would be in my absence.[220] In my absence, they would be slaves to the crueler slave owner, and thus worse off.

Consequently, the plausibility of GP as an appropriate foundation for pre-contractual personal rights is weakened by its allowing one to:

---

[218] Note though, that it is quite possible that I may not have benefited psychologically. But even if we assume that I do, the psychic thrill that I get from seeing Mabel prosper and rise to opulence may not have to do with any material gain that may accrue to me via my friendship with her, nor does my worsening your condition (the downstream fisher folk) need be occasioned by any of Hobbes' dispositions of competition, diffidence and glory.

[219] Donald Hubin and Mark B. Lambeth, "*Providing for Rights*," in *Contractarianism and Rational Choice,* p.114.

[220] Ibid, p.116.

(i) worsen the condition of others very considerably, when doing so is necessary to prevent one's own condition from being worsened only slightly.

(ii) embark on acts of depredation towards others (kill, beat, or rob others) when someone else would do so if one didn't.

(iii) use others in all sorts of atrocious ways as long as in doing this one also helps them in various ways so that they benefit slightly, namely, the net effect of one's interaction with them is positive or the expected utility they get is greater than what they would get in one's absence.

Gauthier might agree with (i) but most likely reject (ii) and (iii) on the ground that they misstate the proviso. Since the proviso constraints interactions only with a view to cooperation, those engaged in predatory and atrocious acts have no reason to envisage cooperation with their victims. Moreover, since the constraints that the proviso provides aim at arriving at mutually acceptable starting point, it will prohibit predatory and atrocious acts since these would undermine the condition of impartiality and equal rationality. Even if we accept as valid Gauthier's line of defense, there is still the question of whether the emphasis of the proviso on material goods does not ignore the value people have "in" or "concerning" themselves as autonomous agents.

3.1.3.2 The Proviso, Bequest, Material, and Non-material Goods

To illustrate the problem that the proviso faces vis-à-vis its focus on material goods, consider the example of Amy[221] and Ben who both live off land that was

[221] Amy's appropriation and transformation of the land could be likened to Eve's. See p.83 and fn 137 as well as pp. 279-280, 290-292 of MbA for Gauthier's discussion of Eve.

initially in common use.[222] Amy appropriates so much of the land. She does many things to the land which thereby increase her level of productivity and wellbeing. She develops it, plants an orange grove, some blueberry bushes, and uses it for bits of manufacturing. But in order not to violate the proviso, i.e. make Ben worse off, she offers him a wage to work on her land. Although Amy's share is more than Ben's, his share exceeds what he was originally producing on his own when the land was commonly held. Amy's appropriation satisfies the first sense or strategy of the six senses or strategies of bettering and worsening above according to which X betters her condition by bettering Y's condition.

Since Amy's appropriation satisfies the proviso, should Ben accept the deal? According to NP and GP, Ben ought to accept it since his condition is not worsened and rather was significantly improved by Amy's appropriation. Amy does well to improve his situation, because the proviso does not require her to do so. What it only requires is that she not worsen his condition relative to the pre-appropriation stage.

It is true that Amy has not worsened the material condition of Ben. It is true also that her appropriation improved his condition. But it does not follow from this that Ben's condition has not been made worse. Wasn't Ben's autonomy violated by the appropriation? Did Amy not use Ben as a "mere means" to her desired end by appropriating the land? Since people generally employ different means to fulfill their diverse conceptions of the good there will be obvious limits to the sort of sacrifice that they can make. Employing people or asking them to sacrifice

[222] Will Kymlicka employs this example in his discussion of the proviso, and my discussion draws on some of his. See Kymlicka *Contemporary Political Philosophy: An Introduction*, second edition, Oxford, Oxford University Press, 2002, pp.116-121.

themselves merely for the benefits of others (without their consent) would be to instrumentalize them. But is this not exactly what Amy did to Ben? Ben did not give any permission to Amy to appropriate part of the land. And he might very well refuse should consent be required to legitimize her appropriation.

In any case, both NP and GP are mistaken to characterize worsening only in terms of material good and welfare. It is true that Ben's material welfare was improved by Amy's appropriation, but why should that be taken to be more important than, for example, the non-material good of autonomy, i.e.—the ability of Ben to act on his conception of himself without being subordinated to the will of Amy? How is Amy to tell that her appropriation does not make Ben non-materially worse off? Or that Ben was not happier in the pre-appropriation stage because of being a joint owner of the land that Amy appropriated?

Notwithstanding his material improvement, Ben might see himself worse off with Amy's appropriation of the land. This will be the case if he thinks the appropriation decreases his autonomy, an autonomy that might be connected to his holistic view of the land and what it stands for.[223] Part of this autonomy might relate to Ben's conception of himself as, for example, a tree hugger or a deep ecologist who sees the preservation of ecosystems, processes in nature, and species

[223] Kymlicka puts it nicely, "But notice that the fact that Ben is now subject to Amy's decisions is not considered by Nozick in assessing the fairness of the appropriation. In fact, Amy's appropriation deprives Ben of two important freedoms: (a) he has no say over the status of the land he had been utilizing—Amy unilaterally appropriates it without asking or receiving Ben's consent; (b) Ben has no say over how his labour will be expended. He must accept Amy's conditions of employment, since he will die otherwise, and so he must relinquish control over how he spends much of his time. Before the appropriation, he may have had a conception of himself as a shepherd living in harmony with nature. Now he must abandon those pursuits, and instead obey Amy's command, which might involve activities that exploit nature. Given these effects, Ben may be made worse off by Amy's appropriating the land, even though it leads to a small increase in his material income," Ibid, pp.116, 117.

as part of his identity or conception of the good. This would have been violated by Amy's appropriation and transformation of the land.

A further problem with NP and GP or the proviso in general is that it violates the assumptions of individual utility-maximizing rationality and mutual unconcern that underlie not just the proviso but interactions in general. Suppose that an individual (let us call him Frank) lives in a community of landowners, all of whom, according to Gauthier "benefit from the stimulus to production that arises from private holdings, and from the limited specialization and resulting exchanges that their holdings make possible."[224] Frank is so charismatic, Gauthier says, he is able to charm all the landowners to bequeath him all their lands with the promise of a reward (perhaps a good life) in the hereafter or heaven. This bequest, i.e. Frank's land monopoly or right of exclusive holding to all the arable land of the community, Gauthier claims, violates the proviso. Why is that so? Because it leaves others in the position of landless laborers,[225] for the right of bequest "must not be so extensive as to afford any individual an inheritance that she could not acquire in some other way without restricting the opportunities or reducing the well-being of her fellows."[226]

We would recall that in specifying the theory of justice in acquisition and the theory of justice in transfer through the proviso, Nozick intends NP to prohibit cases where one either appropriates all the total supply of something necessary for

---

[224] Gauthier, MbA, p.302.
[225] The same sort of reasoning seems to inform Gauthier's argument against "monopolist Eve's" right to control the only oil well on an island. Gauthier's argument is that Eve, who discovers oil on the land she previously appropriated, is not made worse off if her right were rejected although she may worsen the situations of others by being an oil monopolist. Frank is in some respects like Eve, they are at least both monopolists, or potential monopolists. See MbA, pp. 279-280, 290-292.
[226] Ibid, p.302.

life or transfers all holdings into an agglomeration. Needless to say, then, that for Nozick, just as it is for Gauthier, the bequest or transfer of all the arable land (something necessary for life) to Frank violates the proviso. Therefore, Frank cannot claim exclusive right to all the arable land in the community.

This is indeed a curious argument to make. The force of the argument is anchored on the fact that the person who possesses all the arable land or something essential for life reduces the well-being of others. But why should such considerations come into the calculations of rational agents that Gauthier says are utility maximizers and nontuistic? At any rate, it is difficult to see the way in which bequeathing or transferring all the arable land to Frank violates NP and GP since the process of bequest or transfer is done through interaction involving Frank and the previous landowners. Suppose the transfer is negotiated not through a charming charismatic deception of Frank, but in return for a favor that Frank had provided to the landowners or as some sort of reward because they think he is a good person. In which case transferring their individual lands to him would satisfy the requirement of free exchange and the basic assumptions of individual utility-maximizing rationality and nontuistic interests.

Yet, NP and GP would require that Frank give up the right of exclusive holding of all the arable land.[227] To require that Frank turn over the lands that were transferred by people who had prior legitimate claim to them seems to be a great violation of the rights of transfer and acquisition. Beyond this, since the landowners died without leaving behind any family, there is the side issue of who

[227] Because according to the assumption underlying both versions of the proviso the bequest or transfer violates the proviso's requirement of appropriate and legitimate appropriation.

should take possession of the land that Frank is supposed to relinquish. I suspect that Gauthier would say that Frank should turn the land back to the community or that the land ought to remain in an unowned condition, which will be the same as turning it back to the community.[228]

To say that Frank's land monopoly violates the proviso requires that Frank rejects some (if not all) of the land transferred to him. Does this not violate the foundation of free exchange in the market? Should not activities prior to society respect the same principle expected in society? In discussing an essentially just society, Gauthier argues that such a society "can neither ban nor require capitalist acts among consenting adults."[229] If this is so, should we not expect that any state of affairs whether social or presocial that aims towards essential justice to neither ban nor require capitalist acts among consenting adults. But is this not exactly what the proviso does to Frank and the landowners, that is, it prohibits them from engaging in some form of capitalist acts—acts of transfer and exchange?

But suppose that Frank's favor to the landowners was done over a period of time and individually to the landowners. All of the landowners are unaware that each landowner had made a similar bequest and transfer to Frank. Frank, too, although has received a firm promise from the landowners that his favor would be reciprocated in their will is unaware that each of the other landowners has bequeathed their land to him. Over a period, each landowner dies and only then does Frank realize what has been bequeathed to him. But as it happens, the landowners, who constitute 5% of the population of the community, die without

[228] See MbA, p.300 for Gauthier's take on what should be done to an inheritance that was not disposed before one's death.
[229] Ibid, p.341.

marrying and having children. Given that Frank is a very productive person, he puts in as much effort as possible to develop each piece of land as soon as it is passed on to him. After the death of the last landowner, Frank realizes that he now has all the arable land in the community. Meanwhile, he has productively developed and transformed most of the land. The part of the community that capitalist Frank had developed is now home to these infrastructure: schools, hospitals, movie theaters, factories, museums and art galleries, gyms and recreational centers, parks, malls and community centers—all of which now provide entertainment and employment for a significant number of the people in the community. We might even assume that Frank does what Eve does to the land she has appropriated. He "provides effective opportunities" for others to "increase their wellbeing by changing their way of life." On some part of the land he acquires, he introduces new techniques of farming and food production, "increasing yield to such an extent that some of the others may cease to be self-sufficient, becoming instead specialized artisans and crafts persons, exchanging their products for his surplus food and living better in consequence."[230] Given what Gauthier says about efficiency in the state of nature, he is committed to agreeing with Frank that he has a legitimate claim to his acquisition and that the right of bequest satisfies the proviso. He says, "In the state of nature, if not always in society, efficient use is a condition of rightful possession."[231] By increasing the opportunities and wellbeing of others, Frank demonstrates an efficient use of all the arable land and so meets the proviso's condition of efficiency.

[230] Ibid, p.290.
[231] Ibid, p.293

We might wonder, though, why the condition of efficiency should determine the right of bequest or circumscribed and constrain what the landowners and Frank can do with their property. The argument for efficiency does not seem to be available to Gauthier. Gauthier who severely criticized Rawls for his "lexical difference principle" hinges his criticism on the fact that the principle, among other things, licenses freeriding and violates the thesis of individualism by allowing the more fortunate in the "natural lottery" to use as mere means the less fortunate. He writes:

> It appears that the lexical difference principle licenses those with lesser natural talents to take advantage of those naturally more fortunate, requiring the latter to use their abilities, not primarily for their own wellbeing, but to maximize the minimum level of wellbeing. …[Rawls' rejection of personal ownership of natural endowment] leads him to a very different view of what justice requires. The person who takes advantage of her fellows is not the less talented, but the more talented individual who uses her talents solely for her own benefit.[232]

Gauthier's rejection of Frank's right to bequest seems to suffer for the same reasons that he thinks handicap Rawls' lexical difference principle. Both views seem to require what Gauthier himself calls "a very strong condition of mutual benefit,"[233] for "each person's title to benefit [Frank's right to bequest is] dependent entirely on the effect that his receipt of the benefit may have on what others [landless people] receive."[234]

---

[232] Ibid, p.252.
[233] Ibid, p.248.
[234] Ibid, p.248.

I suspect that the intuition that underlines GP's requirement that Frank give up the right of exclusive holding of all the arable land is the same intuition that underlies Locke's position on appropriation. Appropriation, for Locke, must respect the two basic obligations of self-preservation and the preservation of humankind in general. Even though your preservation takes justifiable precedence over that of others in situations where the choice of survival or plundering is between yours and others,' you are not permitted to pursue your interest and self-preservation in ways that leave you "super better" off while endangering the preservation of humankind in general.

In Gauthier's language, all monopolistic activities must be proscribed, for neither Frank nor anyone else is permitted to own an extensive conglomeration of land (through bequest, and probably through some other means) in ways that restrict "the opportunities or [reduce] the well-being" of others.[235] In other words, no one is allowed to become super rich by appropriation that leaves others destitute, or in sufficient need, or in arrant dependence. This is what Frank does by appropriating all the arable land. But is this line of reasoning available to Gauthier?

Locke can very well argue against Frank's appropriation or against any sorts of monopolistic activities because embedded in his theory of rights and appropriation is the view that natural resources and endowments are provided by God for the betterment of humankind in general, hence no one person is permitted to become super rich by an appropriation that leaves others in destitution. God gave us land and all natural resources, which are to be employed for the good of humankind in general. No one person or group of persons is allowed to corner all

[235] Ibid, p.302.

of the land and all natural resources that God gave to all humankind, particularly if such appropriation leaves that person or group super well-off at the expense of impoverishing others, or leaves them destitute, or in sufficient need, or in arrant dependence. Gauthier, however, cannot use Locke's reasoning, for God plays no role in his theory.

3.2 Gauthier's Two Individuals: Economic and Liberal 'Men'

In relation to human interactions, there are two perspectives from which the moral constraints of *Morals by Agreement* can be examined. The first perspective is the morality of the economic individual. For her, moral constraints are viewed as instruments of domination and necessary evil. The second perceptive, which relates morality with an essentially just society is the morality of the liberal individual. Morality, for her, is not an instrument of domination but an indispensable part of cooperative activities. She necessarily occupies the Archimedean standpoint in her choices and actions.

The economic individual is mutually unconcerned and displays a non-tuistic interest in fellow participants. Morality engages not her affections but her rationality. Thus, she is unconcerned about the interest of others while pursuing her own. She finds a lessened instrumental value in morality and takes no interest in constraints derived from the interests of others, except insofar as "adherence to them is instrumental to [her] asocial concern."[236] The lack of tuism of the economic

[236] Ibid, p.326.

individual deprives her of the "affective capacity for morality" and "the capacity for valuing participation."[237]

In contrast to the economic individual, the liberal individual has a vigorous view of participatory activities and displays an affective capacity for (rational) morality. Her emotions and feelings are engaged by the demands of rationally based constraints. She is motivated to value cooperative activities because she values those who make these activities possible. Unlike the economic individual, she manifests tuistic feelings and bonds. Her possession of a sense of duty is a testament to her display of an affective capacity for morality. It is important to stress that an affective capacity for morality presupposes a prior conception of morality, for an individual, as Gauthier rightly observes, "cannot be moved by a sense of duty or justice unless [that individual] antecedently believes some action to be [her] duty [or justice]."[238] If X is moved to do her duty, X does so because she considers it her duty in the first place.

There is the need to distinguish the "capacity for an affective morality" from the "affective capacity for morality." The former is the type of morality that Hume[239] develops—which I consider in the final section of chapter four—which takes humans as constrained by tuistic interests, contingent emotions or feelings in the pursuit of their concerns. The capacity for an affective morality does not

[237] Ibid, p.327.
[238] Ibid, p.328.
[239] Hume's theory of moral sentiments (of sympathy) is an example of affective moralities. Other moralities that bind in a way quite different from a rational morality would be Nozick's 'morality of responsiveness,' 'the morality of care,' 'the morality of benevolence,' and 'the morality of concern,' or the type of empirical morality defended by Richard Brandt, who appeals to empirical laws of nature, i.e. shared sentiments as the firm foundation for morality. See Richard Brandt's *A Theory of the Right and Good*, Oxford, Clarendon Press, 1979. See also MbA, pp.326-329 for Gauthier's discussion of the different ways both affective and rational moralities bind agents.

presuppose any prior conception of morality. It is simply the capacity to be constrained by concern for others in one's asocial pursuits and it is generally "identified as moral in that it introduces constraints, not in that it motivates one to adhere to constraints that one already recognizes."[240] Conversely, the affective capacity for morality presupposes a prior conception of morality. It introduces no constraints, but disposes one emotionally and motivationally to adhere to constraints previously and independently accepted.

Having realized that human dependence and insufficiency is not an evil and that morality serves to enrich human interactions, the liberal individual comes to value those whom she encounters as fellow participants. Group or team activities generally illustrate this sort of human dependence and insufficiency. We can think of an orchestra, with each member exhibiting his or her uniqueness, as each plays his or her individual musical instruments, yet at the same time depending on others, or is constrained by what others do. In the midst of constraints, each simultaneously recognizes the importance of each member in the overall performance. Individual accomplishment, thus, ultimately depends on the group's success. In recognizing this dependency upon others and the value that other members bring to the end goal or value of the orchestra, each performer values one another by taking an interest in the interest of the other. Each realizes that in performing (cooperating) with others she makes "the most effective use of [her] powers to attain certain ends that would otherwise lie beyond [the] individual capacities" of each performer.[241]

[240] MbA, p328.
[241] Ibid, p.345.

This analysis of the interdependence of liberal individuals is in effect an acknowledgment of human finitude. An all-powerful and all-sufficient being has no need for cooperation and for its virtue, justice. But we are not this sort of being, hence, we may not, according to Gauthier, "sensibly suppose that cooperation, which enhances our limited powers and overcomes our individual insufficiency, is a necessary evil."[242] Certainly, to represent cooperation and so justice as a necessary evil is to view being human as itself evil. But this is not the case. For the modes of human activity are great and no person is capable of realizing all of its possible modes. Since the fullest realization possible for each person is the realization of some mode complementary to the realization of other modes by others, the fully rational actor (the liberal individual) neither construes cooperation as an instrument of domination nor dependency and insufficiency as disadvantageous.

One way to view human interdependence is in terms of scarcity—whether productive or consumptive—which depending on how we look at it could be viewed as a human necessary evil or good. Participation thus becomes the method by which this human necessary good or evil is remedied. The liberal individual realizes this. She knows that "not only can no individual realize all forms of human life in herself, but some forms are not individually realizable."[243] This can be said of an orchestra, of games and of almost all social activities.

What the economic individual lacks, the liberal individual has. The exclusively asocial motivation of the former makes her see morality as purely

[242] Ibid, p.345.
[243] Ibid, p.336.

instrumental and an instrument of domination. She thus lacks the capacity to be truly the just and impartial person. The liberal individual who displays tuistic bonds in the value she places on participatory activities and fellow participants balances the failing on the part of the economic individual. The economic individual makes the transition to the liberal individual and in this transformation sees morality as necessary to the pursuit of her self-regarding interests, and no longer does she seek to dominate and exploit others.

There is a similarity between the transformation of the economic individual to the liberal individual and the transformation of the *primitive individual* to the *citoyen* in Rousseau's account of the general will. In both transformations, the individual (economic individual for Mb(CM)A, and the *primitive individual*, for Rousseau) sets aside the disposition to pursue her interest without regard for others or for cooperation. But the transformation takes place when the 'new person' (the liberal individual for Mb(CM)A, and the *citoyen* for Rousseau) acts not on what primarily satisfies her particular interests but on what is mutually advantageous for everyone. There is a difference, though, concerning the transformed individual in both accounts. The *citoyen* has no individuality. Having been subsumed into the general will her individuality has become identical with it. By contrast, the liberal individual has an individuality. Being instrumentally driven she pursues her interests within the constraints of morality.

However, the value placed on participatory activities by the liberal individual is not solely for their own sake; these activities are instrumentally valued to the degree that they promote her desired end. "In the absence of morally

acceptable benefits," she views social activities as exploitative rather than cooperative, and her fellow participants as "having interests opposed to hers."[244] Because the liberal individual values participatory activities instrumentally, the absence of morally acceptable benefits is a threat to participatory activities and consequently a threat to rational self-interest. When participatory activities are not mutually beneficial, the liberal individual collapses back to the economic individual. Since morality must appeal not only to her feelings but also to her reasons, the absence of morally acceptable benefits affects her "development or continuation of tuistic bonds"[245] that are essential for participatory activities.

I have said that the affective capacity for morality is fostered, according to Gauthier, when the economic individual transforms to the liberal individual. It is important to point out that the transformation is a three-step process. First, there is the recognition (by the economic individual) that the constraints of morality are necessary conditions for impartial and mutually beneficial activities. Second, there is the appreciation (by the economic individual) of the value of participatory and shared activities that satisfy the requirements of these constraints. Finally, there is the cultivation of tuistic interest in fellow participants such that the economic individual affectively values the morality that first appeared only as a rational constraint.

One recurring theme in my analysis of the development of tuistic bonds is that the transformation from economic individual to the liberal individual reflects the value she places on participatory activities and fellow participants. She values

[244] Ibid, p.338
[245] Ibid, p.338.

participatory activities and fellow participants when she does not seek to exploit them for personal gain. She recognizes that society is a cooperative venture for mutual advantage. But if society is a cooperative venture of mutually unconcerned individuals, in what sense is the liberal individual tuistic?

3.2.1 The Liberal Individual, Tuistic Bond, and Free Affectivity

To say that the liberal individual develops tuistic bonds and takes an interest in the interests of fellow participants in the pursuit of her interest (because she recognizes the mutual unwillingness of participants not to take advantage of others) does not mean that she is "altruistic," or that she has the capacity for affective morality. She is neither a candidate for Rousseau's general will nor is she Hume's ideal sympathizer. The interest that she has for others is still instrumental: others are valued because of the value she places on the benefits of participatory activities. By valuing these benefits, she instrumentally values the activities and those that make them possible.

The liberal individual instrumentally values fellow participants and exhibits a cooperative or CM disposition towards them by her willingness to promote an essentially just society. She is motivated to engage not just in fair deals but also in cooperative activities that satisfy the demands of justice and maximizes the utility functions of fellow participants. She recognizes that others are willing to participate in cooperative arrangements insofar as they consider them beneficial and impartial. To the extent that cooperative arrangements are beneficial compared

to universal noncooperation, the liberal individual considers it rational to cooperate.

The point about tuistic bonds relates to free affectivity. The liberal individual is free to form whatever affective ties she likes as long as these ties are not part of the constraints of Mb(CM)A. The constraints of morality bind rationally and independently of all particular affective preferences. The liberal individual makes her choice of others as objects of affection. She is not bound by fixed social roles, either in her activities or in her feelings. Free affectivity recognizes individual autonomy and an essentially just society encourages it within a robust development of the affective capacity for morality. An essentially just society does not impose emotional bonds on people; it allows them to enter voluntarily into enduring and binding relationships with others.

Allowing agents to develop freely an affective capacity for morality is very important for the stability of society. What it does is that it ensures, among other things, that the "constraints required by essential justice" are accepted willingly rather than imposed externally. A society lacking members who have this capacity, in order not to "rapidly destabilize," would have to impose "on them, through processes of socialization, loyalty to a more substantive goal, which would define roles that individuals would not be free to accept or reject."[246] But a society that imposes this capacity and does not allow individuals to enter voluntarily into enduring ties and binding relationships with others treats them as if they were children, and as means to some goal or conception of the good that holds true independently of them.

[246] Ibid, p.348.

Closely related to the idea that the liberal individual is free to form whatever affective ties they consider appropriate is the idea of an independent conception of the good. The liberal individual has her own independent conception of the good. Being fully rational—where "rationality embraces both autonomy and the capacity to choose among possible actions"[247] on the basis of an individual's conception of the good as determined by that individual's preferences—the liberal individual is aware of the reflective process by which her later self emerges from her present self, so that her preferences are modified not in a random or uncontrolled way but in the light of her own experiences and understanding. This point about the link between present and future selves and preferences raises the interesting issue about the uncertainties underlying the liberal individual's pursuit of her interest.

3.2.2 The Liberal Individual and Uncertainties Underlying Pursuit of Interest

Unlike the economic individual, the liberal individual recognizes two primary uncertainties that underlie her pursuit of interest, namely, her utilities. First, there is the uncertainty that relates to the concern of stability, which she recognizes could be addressed by engaging with others in ways that elicit their willing and rational cooperation. The willingness of others to voluntarily accept social institutions and practices is predicated on the fairness of these social institutions and practices. Hence, voluntary compliance is possible only in an essentially just society. In an essentially just society, institutions and practices accommodate themselves to all persons, whatever their preferences and capacities are, provided of course that they

[247] Ibid, p.346.

are participants in cooperative activities. Unlike Rawls' individuals whose preferences are contoured by social institutions and practices, the preferences of Gauthier's individuals does the contouring; they effectively determine the direction of social institutions and practices.

The second uncertainty concerns the issue of the relationship between temporally distant selves. More particularly, the issue concerns the "reflective dimension of rationality" as it relates to the issue of revisability.[248] For the self at time t_1 could very well have a preference for treating certain cultural goods or certain features of nature as valuable and insist that they factor in decisions setting up cooperative society. But this same person at time t_2 may discount these goods and features or even hold preferences that exclude them for consideration altogether. The problem then becomes how to relate the self that values *v* and that has certain preferences at time t_1 to the self who values *w* and has different preferences at time t_2? Gauthier's take on this issue is that both selves are the same to the extent that the expected preferences of the latter self are given significant weight in the preferences of the earlier self. He writes:

> "The self at time t_1 is identical with the later self at time t_2 to the extent that it identifies with the later self, and this identification is measured by the weight given to the expected preferences of the self at t_2 in the preference of the self at t_1."[249]

Gauthier's response suggests that there is a correlation between the reflective dimension of rationality vis-à-vis the principle of revisability and the

[248] Ibid, p.342.
[249] Ibid, p.343.

exercise of rational freedom. This correlation, according to Gauthier, gives rise to prudential base preferences. Furthermore, it helps us to understand how two selves at times t_1 and t_2 could be related. He states:

> The person who is concerned with the full exercise of his rational freedom cannot agree to social institutions and practices that are merely instrumentally just, however well adapted they may be to his present concerns and powers, because he has no guarantee that such a social structure will continue to provide fairly for his satisfaction.[250]

There is a tension lurking in Gauthier's views of the self, preferences and revisability, which he seems to be unaware of. Cleary, there are issues that anyone dealing with the self and personal identity must grapple with. Issues like what am I? Am I circumscribed by psychological or physical continuity, or by memory, or by a soul? When did I begin? What will happen to me when I die? These, however, are not the sorts of issues that concern us. The issue we are concerned with is how Gauthier's view of the self addresses the worry about massive changes in circumstances and technology. For instance, a world of huge technological change, where, for example, other things are constant would pose a challenge to the self. A self that is finite may find it out of range to give weight to future preferences or form at time t_1 futuristic preferences that she may hold at time t_2. Unable to foretell the trajectory of the massive technological modification at time t_2 she would be unable to incorporate her preferences at time t_2 in her time t_1 preferences.

Because the individual cannot forecast what future technological changes would come about, she cannot reasonably predict whether her utilities would be

[250] Ibid, p.344.

maximized at time t_2 and, hence, cannot have at time t_1 well-defined preferences of the social scheme available at time t_2, and as such cannot give adequate weight to them. Given this dimension of Gauthier's theory, the best she can hope for and which she might reflect in the preferences of the earlier self is a social scheme at time t_2 that she expects is better than a social scheme at time t_1 or one that she expects would sufficiently maximize her utility profile. Because this expectation is at best sketchy and indeterminate, preferences at time t_1 that aim to incorporate preferences at time t_2 would, if anything, be incomplete, if not completely vacuous.

Gauthier could run away from this difficulty by arguing that even though the self is finite, The individual has more capacities to project positively into the future than I have ascribed to her, especially when the issue that is being considered is her wellbeing and utilities. However, I do not believe this tack quashes the problem, since there are limits to what any finite self can predict and know about the future. There is a better alternative route available for Gauthier in dissolving the tension. He can parse utilities into two kinds, namely, those that are EU-focused and those that go beyond them, i.e. value-oriented utilities or utilities that are about the meaning of actions.

Call the first 'the utilities of outcomes' or outcome-sensitive utilities and the second 'the principle' (or 'utilities') of actions, or action-sensitive utilities. The former utilities are attached to outcomes and the latter to actions. The utilities of actions provide us a way of cashing preferences such that whatever preferences or indeed substantial alteration in technology occurs at time t_2, these would already have been anticipated in the individual's preferences at time t_1. This would be in

virtue of the fact that certain actions and their likes at any time or in the n^{th} time express something that the individual holds at time t_1 or, as we discussed in the last section, principle (P) is expressed by actions (A).

We might say, for example, that taking off my shoes when I visit a home expresses some principle or value about politeness or good manners. Other like actions or classes of actions could very well still express this same principle or value of politeness, such that at any time I would always choose these actions because I am a polite individual who at all times is keen on seeking out actions that express this principle. At time t_1 I always do A, namely, I always take off my shoes because this act represents P, i.e. politeness. Now, at time t_2, due to technological changes we neither have shoes nor wear them—we walk barefoot. If we do not wear shoes at time t_2 I would be unable to express politeness by the particular act of taking off my shoes when visiting.[251] However, the fact that there are no shoes for me to take off at time t_2 does not mean that I cannot express P.

Given that like actions or classes of actions, call it X-like actions, express politeness, we could say that what is not available to me at time t_2 is only a particular action. Specifically, what I don't have at time t_2 is not the entire class of action that express politeness, but a particular action, the action of taking off my shoes. Shoe taking-off would then be an instance of the class of polite actions at time t_1 while other like actions would then be instances of polite actions at time

[251] I am not sure whether 'a world of shoe wearing' is superior to 'a world of walking barefoot.' It might very well be, but this in no way affects the discussion about massive technological changes. Technological alterations could be progressive or retrogressive. We could very well move from our present state of industrial development to one that is largely agrarian or to one that is more sophisticated. We could move from one of affluence to one of extreme scarcity and poverty. Changes or movement from time t_1 to time t_2 could either be positive, or negative, or neither.

t_2.We may suppose that belonging to the category of polite actions at time t_2 are like actions such as wiping the dust off one's feet when visiting a home, identifying others by certain titles when talking to them, or looking at them in certain ways when conversing with them. If these actions adequately confer deference to others, they would fall into the category of polite actions at time t_2 and by choosing any of them I would be expressing P.

We may call the action of shoe taking-off at time t_1 *X^1 action*, and those actions at time t_2—wiping the dust off one's feet as a guest, or identifying people by certain titles when talking to them, or any other action that sufficiently confers deference to others—*X^2 actions*. From this, we can say that both X^1 and X^2 actions belong to the class of *X-like actions*—for each set of actions is an instance of Xness—such that once the class of X-like actions satisfies P, any action that is in the class like those of X^1 and X^2 satisfies P as well. Therefore, my doing any of the action that falls into this class at any time (t_1, $t_{2...}$ n^{th} time) sufficiently expresses politeness. So even though we do not wear shoes at time t_2, we do have other acts that are 'like wearing shoes' in the sense that they represent politeness.

3.3 What is it that we have Achieved thus far? How is Mb(CM)A Faring, and Where are we Going?

It is clear from my analysis of Mb(CM)A that Gauthier has indeed devised a rigorous, systematic and strikingly sophisticated account of morality as an addition to the contract tradition. Mb(CM)A is to date by far the most persuasive and insightful approach in this great tradition. In linking rationality to morality,

Gauthier makes an exceptionally good case that morality can be individually rational. The idea underlying this conclusion is intuitively appealing: morality *must* speak to our reason if it is to constrain our behavior and to motivate us to action.

By making a very good case for morality within reason, Mb(CM)A does undercut views that found the moral enterprise on irrational moral motivation, external coercion or eternal conceptual relations that hold true independently of our rational self-interest, namely, considered coherent preferences. But is Mb(CM)A 'capable of withstanding critical examination?' For the most part it does, even though it succumbs to some serious objections, one of which is the problem of secession, which I shall discuss in chapter four. Some of the objections to which Mb(CM)A succumbs to are evident in this chapter. Let me by way of summary examine two of them.

First is the issue of the rationality of CM or the disposition to cooperate. In my analysis of CM, it was obvious that Gauthier is right that it is one of if not, the most fruitful idea or sub-theory in Mb(CM)A. By identifying rationality with utility-maximization at the level of dispositions to choose, Gauthier steers the individual away from limited cooperative activities and activities that offer less than optimal outcomes. By attending to dispositions rather than to particular actions or mere behavior, Gauthier was able to explain why it is rational to perform those actions which are demanded by morality even though they do not directly maximize expected utility.

But it was clear in my examination that there were tensions in Gauthier's argument for the rationality of dispositions. Most of the pitch I made for CM and

the disposition to cooperate seems to assume (a) that dispositions have some quasi-magical property, i.e. that adopting a particular disposition will cause the individual who forms the disposition to act on the chosen disposition and others to adopt a similar disposition, and (b) that disposition is a causal mechanism that cannot break down. This seems quite implausible. Despite the transitivity argument that P = D, D = A, therefore P = A, the appeal made to the causal efficacy thesis seems to render CM as a rational strategy relatively dubious and at worst unrealistic. For if the thesis were true, it is hard to see how the PD could have been the deep social and historical problem for social cooperation that is has been for many years.

That the jury is still out on the question of whether CM is a rational strategy for individuals to adopt is quite evident in my response to the claim by Copp that reserved maximization is a better strategy than CM. It is quite apparent that my response did not completely address the core of his argument that reserved maximization yields greater utilities than CM. The core of Copp's objection is that if the reserved maximizer is not reasonably sure he would be able to get away with violating a requirement of a scheme of cooperation whenever he had the opportunity to win a jackpot, he would not do so. But if he is absolutely certain that he can get away with it he would violate a requirement of cooperative scheme whenever he had the opportunity to win a jackpot. The reserved maximizer accepts the argument that constraints are essential to any scheme of cooperation, but he is prepared to take the 'money from a lost wallet, provided enough money is involved,' and provided he is absolutely convinced he would not be caught. He is prepared to steal his neighbor's goods, provided there are of great value to him, and

provided he knows he can get away with it. The point is that in these situations, the reserved maximizer has no rational motivation, on Gauthier's theory, for not stealing either the money from the wallet or his neighbor's goods.

Mb(DV)A replies to this sort of objection and similar ones, in a way that Mb(CM)A is not able to. When EU is stacked too high against cooperation or not stealing the money or the goods of one's neighbor it is not rational to cooperate or not to steal the money or the goods. Mb(CM)A succumbs to these sorts of objections unless it is interpreted as I am proposing, namely, along the direction of Mb(DV)A, in which case the most serious objections fall away. Mb(DV)A does not succumb to these sorts of objections because in jackpot situations or in situations where EU is stacked too high against cooperation, it may or may not be rational to steal either the money from the wallet or the neighbor's goods. Whether it is rational for an individual to steal the money or goods or whether it is rational for an individual to cooperate or not would, according to Mb(DV)A, depends upon what each of the act means for that individual, in addition to their possible outcomes.

Now to the second issue: the issue of constraining appropriation in natural interaction, i.e. in the pre-bargain stage by the proviso. Gauthier rightly recognizes that issues of impartiality and stability would affect the contract if the initial position is not constrained by the proviso. But was he successful in arguing that the pre-bargain position is based on constraint from the proviso? It was obvious from my critical discussion of the proviso that it is both too strong and too weak. How is that so?

It is too weak because it allows us to worsen the situation of another person when interacting with that person for the benefit of a third party, provided we do not benefit from such interaction. Also, it permits us to worsen an individual non-materially, provided we improve their material condition or do not leave them worse off materially. The proviso is too strong because it prohibits certain types of appropriations, bequests, or transfer of property. The proviso forbids appropriations or bequests that make an individual or a group of people considerably better-off but leaves others considerably less well-off even when it is clear that these activities satisfy the requirement of free exchange and transfer. Specifically, the proviso requires a very strong condition of mutual benefit, whereby a person's 'title to a benefit is dependent entirely on the effect that the person's receipt of such benefit may have on what others receive.'

Mb(CM)A's view about appropriation, bequest and transfer is exploited by Mb(DV)A. Mb(DV)A sides with Frank that his claim to all the arable land is legitimate. It recognizes, however, that Frank might give up some (if not all) of the land that was bequeathed to him. By giving up his right to the appropriation of all the arable land in the community, Frank increases or maximizes the expected utility of others. However, in doing this, his expected utility is diminished. If this violates his right of legitimate acquisition or the foundation of free exchange in the market, Frank would have no reason to give up his right of exclusive holding of all the arable land, except for the fact, of course, that he believes that his doing so provides him utilities outside those (i.e. EU) that he gave up. Because EU is stacked too high against Frank's giving up the land, he has no EU-reason on

Mb(CM)A EU-focused account to give up some (if not all) of the land that was bequeathed to him.

Mb(DV)A claims that Frank may have other reasons and may be rational to give up some (if not all) of his claim to the arable land that was bequeathed to him. For example, if Frank believes that acting in ways that reduce the well-being of others has some negative value for him, say it symbolizes for him unkindness or malevolence, or if giving up some (if not all) of his entitlement to the arable land expresses for him benevolence or generosity, we would expect him to factor this into his decision whether he should keep or give up some (or all) of the arable land in the community. If he is reluctant to acquire all the arable land or if he is willing to give some of the land back to the community we can understand his reluctance or willingness not merely as an indication of his belief that by keeping all the land to himself he lessens the well-being of others in the community—this might be it too—, but rather as an indication that the actions (i.e. not acquiring all the arable land or giving some of the land back) speak to some deep value that he believes in. This would be quite apparent when I discuss fully Mb(DV)A in chapter five. For the meantime, I turn my attention, in the next chapter, to one of the major problem confronting Mb(CM)A—the problem of secession.

Chapter Four

The Problem of Secession and Moral Theorizing

Introduction

The problem of secession is a test of application of Gauthier's brand of contractarianism. There is a great deal in common between this problem and another problem that Gauthier discusses in an article he wrote in 1978: the scope of the contract problem. Common to both problems are issues concerning the nature, scope and parties to agreements. The fundamental questions that arise in connection with these problems are, "What is it that makes an individual eligible for contract-membership? And who ought to benefit from the gains produced by schemes of cooperation?" In this chapter, I shall be examining the problem of secession. My examination of this problem proceeds from my analysis of the scope of the contract problem. I shall be demonstrating how the narrow and misleading characterization of rationality by Gauthier's moral contractarianism informs its solution to both problems in general and the problem of secession in particular. I hope to be able to show from my examination of the problem of secession and Mb(CM)A's solution to it, that the theory provides a single-tracked silver bullet solution to a problem that requires a multi-tracked framework for solutions. In concluding the chapter, I shall discuss the extent to which Hume's theory of moral sentiments justifies Braybrooke's positive thesis that an account of affective morality or a theory of moral sentiments dissolves the problem of secession.

4.1 JaF, the Thesis of Individualism, and the Problem of Secession

In chapter two, I examined Rawls' contribution to the social contract tradition. My analysis of JaF primarily focused on one of his two special formulations of the general conception of justice: the lexical difference principle, the other being the liberty principle. The difference principle states that social and economic inequalities "are to be arranged so that they are both to the greatest benefit of the least advantaged."[252] I noted that it is an egalitarian principle because it supports a system that redistributes opportunities, resources and benefits. Since the principle takes social practices and institutions to be just and fair to the extent they benefit the less endowed or least favored members of society, it constitutes part of the overarching framework required, according to Rawls, to achieve and maintain the bases of self-respect that is crucial to upholding the standard of equality and fairness.

The difference principle considered along with all the assumptions in JaF, i.e. the veil of ignorance, the maximin rule, indeterminate persons—possessing a capacity for a sense of justice—and the method of reflective equilibrium represent an elegant and abstract version of contractualism. I claimed that if we accept the method of reflective equilibrium and all the assumptions, then Rawls' version of contractualism constitutes a persuasive and fruitful approach to distributive social justice and to the great contract tradition. As long as the assumptions that Rawls makes take their rightful places within the framework of his brand of contractualism Rawls can rely on them to do their job; and they do a very good job in strengthening

[252] Rawls, ToJ, p.386.

his overall argument that agents would willingly comply with agreements that they previously entered into.

As far as the general problem of rational compliance and the problem of secession are concerned, JaF makes a good case that agents will choose to cooperate by following through on terms of agreements.[253] When it comes to both problems or to the questions that motivate both problems—"What rational motivation do I have for not going back on my word when it seems advantageous for me to do so? What rational motivation can I have for not joining others that are better-off like me in forming a society of well-off members?"—JaF's response seems unequivocal. Given that (a) a society that is structured by the two principles of justice is *the* most equitable and just, and (b) I possess the capacity for a sense of justice—which is one of moral powers that define me as a moral and autonomous person—I have a rational motivation to honor agreements and to cooperate with others. For if the principles of justice are chosen from the standpoint of the least advantaged member of society, then no matter where an individual ends up in the social and economic spectrum that individual would benefit. Furthermore, given

[253] Braybrooke has argued that unlike Gauthier's moral contractarianism Rawls' theory fails to solve the general problem of rational compliance. He writes: "[I]n principle Gauthier has solved, so far as rational argument can solve, the compliance problem that remains in Rawls's argument for agents who, after the veil is lifted, find that they are not in the least-advantaged stratum. Every one of Gauthier's agents, winners or losers in character and skills, has grounds sensitive to his individuality for accepting the scheme that justice prescribes because every one of them was represented, individually, in the choosing of the scheme. The other compliance problem, which Rawls simply assumed away, of acting in accordance with the scheme indefinitely, once it has been accepted, Gauthier solves by arguing that the contracting parties will as rational agents induce in themselves a reliable disposition to comply whenever they have to deal with other agents that they find transparent enough-translucent-to be relied on to comply in turn." "Social Contract Theory's Fanciest Flight," in Ethics. I agree with Braybrooke, but I add that if JaF fails to solve the general problem of rational compliance—and the problem of secession—it fails mostly because the assumptions that Rawls appeals to are rejected. Therefore, if JaF seems to fare better than Gauthier's moral contractarianism when it comes to either problem it fares better because its many assumptions are accepted, and if it seems to fare worse it is because there are reasons convincing enough for not accepting the assumptions.

that agents can be relied upon to comply with the principles of justice, since they possess a capacity for a sense of justice there is no reason to think that they would be moved to secede from the larger society because they would have benefited more if some other principles were chosen.

However, accepting JaF comes with a big price, namely, as a theory of distributive social justice it violates the thesis of individualism—which stipulates that we may neither collapse a person's conception of the good with those of others, nor compel anyone to accept the principles of social relationships—that is fundamental and important to liberalism. It is argued that persons are individuals and are individuated by their conceptions of the good. Since they pursue their various conceptions of the good independent of one another there is a limit to the sorts of sacrifice that one person can make for the benefits of others. In the face of the criticism, Rawls maintains that the difference principle provides the means and enabling resources for people to pursue their various conceptions of the good, and by providing these means and resources it in fact respects persons as ends in themselves. The principle ensures that no one depends on others for the protection of their interests and that all are socially and economically self-sufficient. Because persons *qua* citizens are not subservient to the will of others they can respect one another as equals and not as superiors and subordinates.

Gauthier has argued that the difference principle is implicitly collectivistic at its worst, and like aggregative theories, namely, average-utilitarian principle[254] obscures the strong sense of individualism. By obscuring the strong sense of

[254]Average-utilitarianism is different from classical utilitarianism. The former prescribes maximization of average utility or happiness, while the latter prescribes maximization of total utility.

individualism underlying social contract theory, the lexical difference principle betrays Rawls' avowed commitment to the idea that rational agents are separate and equal decision-makers in choice situations. To consider the contract as an individual decision based on the perspective of the least advantaged member of society, Gauthier argues, is to impose a strong egalitarian constraint on the pursuit of individual interests in society, according to which each person "may benefit only on terms which maximize the minimum benefit."[255] How does Gauthier's brand of contractarianism fare in the test of application? It is this question that I turn my attention to.

4.2 Mb(CM)A, the Scope of the Contract Problem, and the Problem of Secession

It was Gregory Kavka who coined the phrase, the problem of secession in reference to Gauthier's worry of the scope of the social contract .[256] The problem of secession is a problem that arises within the life of the contract. It is a problem concerning what ought to be done with previously productive members of society who have for some reason become unproductive. The question that arises in connection with the problem is this: is it rational for an entity (a subgroup or a group of people) to secede from society (or a larger group) just in case agreements or that society fail to satisfy the demand of mutual advantage? Stated differently, is it rational for better-off or more-favored members to cooperate with or support less well-off or less-

[255] David Gauthier, "The Social Contract: Individual Decision or Collective Bargain?" in *Foundations and Applications of Decision Theory*, Vol. 2, C.A. Hooker, Jim Leach, and Edward McClennen (eds.) Dordrecht, Holland, D. Reidel, 1978, p. 66.
[256] See Gregory Kavka, *Moral and Political Theory,* Princeton, Princeton University Press, 1986, pp.240-243.

favored members of society when EU is stacked too high against cooperation? This question arises within the larger questions, "What is it that makes an individual eligible for contract-membership? And who ought to benefit from the gains produced by schemes of cooperation?"[257]

The scope of the contract problem concerns the rationale for partitioning humans into groups, groups of nations or nation state. I turn to Gauthier's discussion of the problem.

> No doubt the members of the most existing human groups are all better off than they would be were they in the non-social state of nature. Insofar as this is the case, the members have a basis for cooperative arrangements among themselves, which we have identified with society. It may then be argued that this justifies the existing national divisions. But clearly this is not so. For there are many alternative ways of partitioning mankind into groups such that everyone would be better off than in the state of nature. What we require is a justification of some particular division, which would not serve equally well as justification of other, incompatible divisions.[258]

But why should we take the contract as a "contract of national societies" and not a contract of "the society of the human race?"[259] The reason is that only the former meets, according to Gauthier, the demand of mutual advantage. I say more about this in a moment.

---

[257] On a practical level, we could point to societies where secession issues play out. There is the subtle push or more than subtle push by some Québécois elements for Québec's separation from the Canadian Federation. There is too the growing suggestion of Alberta's 'independence,' propelled, among other things, by economic considerations. There is also the push for secession by some part of the Southern region in Nigeria. This move for secession is primarily driven by the absence of mutual advantage in the Nigerian Federation, i.e. the freeriding on the Southern region by the Northern region. These examples may indeed provide insight into some nuances of issues of secession. My analysis in this work, however, is purely theoretical.
[258] Ibid, p 61.
[259] Ibid, p.61.

The scope of the contract problem is similar to 'the stateless or open border problem' that confronts Rawlsian (egalitarian) liberalism. The open border problem is a problem of how to reconcile liberalism's deeply held principle of moral equality of persons with the existence of separate states.[260] It is true that liberalism claims that persons are morally equal. Also true is that liberalism supports national and group divisions. Does liberalism not violate the liberal principle of moral equality of persons when it supports separate states or prohibits open borders? Some egalitarian liberals (e.g. Will Kymlicka) appeal to 'group-differentiated rights' to defend the liberal position on national societies. The argument for this proceeds as follows.

> **P1:** Cultural membership provides a veritable and rich context for individual choices and the pursuit of an individual's conception of the good.
> **P2:** Individuals or people are members of societal cultures or group memberships.
> **P3:** If we are to provide people a veritable and rich context for choices and pursuit of their conceptions of the good, it is necessary to protect societal cultures.
> **P4:** The existence of the state is necessary for the protection of societal cultures.
> **Conclusion:** Therefore, the state exists both to provide and facilitate the contexts for individual choices and pursuit of conceptions of the good and to protect individual cultures.[261]

---

[260] See Will Kymlicka's "Justice and Minority Rights," in Robert E. Goodin and Philip Pettit (eds.) *Contemporary Political Philosophy: An Anthology*, Oxford, Blackwell, 1991 pp.378-390 for a statement of this problem and how liberals can deal with it by appealing to group-differentiated rights.
[261] Ibid, p.379.

Gauthier does not build his argument for partitioning individuals into groups or national societies around the idea of securing and protecting group-differentiated rights, and understandably so. He builds it rather on mutual advantage, defined in terms of the maximization of expected utility. He argues:

> It is a necessary, but not a sufficient, condition for rational agreement among a group of persons, that each prefers agreement to no agreement. But it is also a necessary condition for rational agreement that there is no subgroup, each of whom prefers agreement only with other members of the subgroup, to agreement with all. *It is not rational for persons to extend their agreement more widely than benefits them.*[262]

What would be the implication for the scope of the contract if the contract is circumscribed by mutual advantage? Let us consider Gauthier's discussion of the two perspectives from which we can evaluate net benefits:

> Suppose that mankind is divided into two groups, the inhabitants of developed countries, and the inhabitants of less-developed countries. Then it is plausible to argue that every person in the developed countries would be better off if the developed countries constituted a contractually-based society, leaving their relations with the rest of mankind in the state of nature, than if there were a single society of human race. And it is even more plausible to argue that every person in the less-developed countries would be better off if there were a single contractually-based society of the

[262] Gauthier, "The Social Contract: Individual Decision or Collective Bargain?" p.62, my emphasis. In MbA, Gauthier claims that the condition for mutual advantage being a necessary condition for the acceptability of a set of social arrangements as a cooperative venture, presupposes the reasoning underlying the feminist thought on what is fundamentally problematic with the core form of human exploitation. The mutual benefit of contractarianism distances itself from this exploitation by insisting "that a society could not command the willing allegiance of a rational person if, without appealing to her feelings for others, it afforded her no expectation of net benefit", p.11.

> human race, than if the less developed countries constituted a society, leaving their relations with the rest of mankind in the state of nature.[263]

On the one hand, we have the first group, the group of the inhabitants of more developed societies or countries. Let us call this group Nation S. Members of Nation S favor national societies because national societies work to their benefits. Given that national societies satisfy mutual advantage, members of Nation S, according to Gauthier, are justified to favor a partitioning along national societies over one that consists of the entire human race. If members of Nation S "prefer to belong only to their national societies, and not also to a society of the human race" on the ground that only the former meets the demand of mutual advantage, then they will, Gauthier says, insist that in the "state of nature, they would rationally have entered a contract, not with all other persons, but only with members" of Nation S.[264] Because it is not rational for persons to extend their agreement more widely than benefits them, it is rational for members of Nation S to insist on national societies just in case such partitioning and not some other, say, a partitioning of human society provides them greatest expected utility.

On the other hand, we have the second group, the inhabitants of less-developed countries. Let us call them Nation N. Members of Nation N do not favor national societies; rather, they favor a society of the human race. We may suppose that members of Nation N, whose preference for a world society is informed by their less developed social and economic status are representative of the worst-off group in JaF, except in this case they have knowledge of their identities,

[263] Ibid, p.62.
[264] Ibid, p.62.

characteristics and circumstances. Like the least advantaged members in JaF they insist that a society of the human race satisfies the condition of mutual advantage because it promotes conditions that make it possible for everyone to purse their conception of the good. Their insistence for the creation of a single membership in a society of the human race, Gauthier argues, cannot be justified on the ground that it promotes conditions that make it possible for everyone to purse their conception of the good. For to require members of Nation S to extend their agreement more widely than benefits them is to require them to produce benefits for others or to engage in sacrifices for the benefits of others. To grant Nation N its demand is to violate the thesis of individualism, as well as the condition of mutual advantage.

Now if mutual advantage justifies, as Gauthier says, the preference for national societies by members of Nation S, then, on the force of it, endowed or better-off contractors, according to Kavka, are justified to secede from a society involving the endowed or productive and unendowed or unproductive. Kavka says, and I quote him at length:

> This raises the specter of more-favored negotiators seceding from the negotiations to form a commonwealth with fewer, but on average more productive, members. Thus, for example, the endowed negotiators might break away and form their own State, leaving the unendowed behind, or the nonhandicapped negotiators might reach an agreement among themselves that excludes the handicapped. Allowing individuals knowledge of their own personal characteristics…opens the door to this possibility that certain subgroups would view it as in their interests to secede from the larger groups and found a State among themselves."[265]

[265] Kavka, pp.240, 241.

Kavka's point is that any theory that both allows individuals knowledge of their own personal characteristics and circumstances, and defines mutual advantage strictly in terms of the maximization of expected utility, as is the case with Mb(CM)A opens the door to secession. Knowing that they are endowed and better-off, and knowing also that supporting a particular scheme of cooperation involving them and the unendowed and less well-off would mean that they have to extend their agreement more widely than benefits them, better-off members will move to secede in order to form a commonwealth among themselves.

The scope of the contract problem we might say is one that arises before the contract. It raises the basic questions, what is it that makes an individual eligible for contract-membership? And who ought to be included in the contract or bargaining process? The problem of secession is one that arises within the life of contract. It raises the fundamental questions, what is it that makes an individual eligible for contract-membership? Who ought to benefit from the gains produced by schemes of cooperation? And is it rational for better-off contractors to extend their agreement to those who do not benefit them, i.e. is it rational for them to support less well-off members even though such act fails to maximize their expected utility? Insofar as the scope of the contract problem and the problem of secession are concerned with issues about the nature, scope and parties to agreements they share a great deal in common. Consider the following possible schemes of cooperation and some of the problems that they generate.

1. A social scheme of individuals not yet joined in cooperation but willing to cooperate, all of whom are variously naturally endowed, i.e. everyone possesses in various level initial factor endowments.
2. A social scheme of individuals not yet joined in cooperation but willing to cooperate, some of whom are naturally endowed, i.e. while some individuals possess initial factor endowments some don't.
3. A social scheme of individuals joined in cooperation, some of whom can no longer contribute to the cooperative surplus, i.e. there are some who are less well-off and may need the support of those who are better-off.

We might call the problem that arises in the first possible scheme of cooperation the bargaining problem. By this, I mean the problem of getting every party to agree on the method or principle of divvying up the cooperative surplus—the sort of problem that confronts money market investors Abel and Mabel. Once contractors settle on the terms of the bargain, including the principle of distribution the contract can take a life of its own. Any problem that arises at this stage generally has do with the problem raised by the rational skeptic, namely, the problem of making sure that anyone that shares from the cooperative surplus continues to contribute to it. If we can get everyone to constrain his or her behavior such that each person is prevented from acquiring benefits from cooperation without paying the necessary costs we would have solved the PD. As we saw in chapter three, Mb(CM)A claims that we can solve the PD if we identify rationality with utility-maximization at the level of dispositions to choose, i.e. along a CM dimension.

The problem for the second and third possible schemes of cooperation is slightly different from that of the first. On both schemes of cooperation, the

problem is about what ought to be done with unproductive and less well-off individuals. In the second possible scheme of cooperation, contractors have not yet settled on the terms of the bargain. Some people are unable to contribute to the cooperative surplus because they have no initial factor endowments. Given that these ones do not possess factor endowments and given that they cannot contribute to the benefits produced by scheme of cooperation should they be included in the contract or allowed to participate in the bargain process? This we might say is the scope of the contract problem. The third possible scheme of cooperation presents us with the problem of secession. In this scheme of cooperation, as it is with the first scheme contractors have settled on the terms of the bargain. Although the contract has taken a life of its own and although it is not bogged down by the problem raised by the rational skeptic, it is confronted by the problem of what ought to be done with unproductive and less well-off members of society. The unproductiveness of those in this scheme of cooperation arises within the life of the contract. They previously contributed to the cooperative surplus but due to some reason they are no longer able to do so. Should agreement be extended to them even though they no longer benefit us?

What is Mb(CM)A's response or solution to the scope of the contract problem, i.e. the problem of the second scheme of cooperation, and the problem of secession, i.e. the problem of the third scheme of cooperation? Mb(CM)A's solution to both problems is suggested by its conception of rationality, according to which it is not rational for individuals to extend their agreement more widely than benefits them. Given this view of rationality, it would be rational for more-favored

contractors to exclude from the contract those who do not possess factor endowments. More-favored contractors neither bargain with those who have no factor endowments nor form a society of the human race, i.e. a society consisting of more-favored and less well-off contractors. Rather, they bargain with other more-favored contractors, i.e. they form a national society made up of more-favored contractors.

Mb(CM)A's solution to the problem of secession is similar to its solution to the scope of the contract problem. Given its conception of rationality, well-off and productive members are better off not cooperating with unproductive members of society. For them to support unproductive members is to extend agreements more widely than benefits them, which in turn is to violate the thesis of individualism as well as the demand of mutual advantage, namely, to breed freeriders. In simple terms, it is not rational, according to Mb(CM)A, for productive and better-off members to cooperate with unproductive and less well-off members. They are better off seceding from the larger society and forming a commonwealth among themselves.

If we sketch a matrix for the problem of the third scheme of cooperation and Mb(CM)A's solution to it, we have the following table. Group S represents the group of better-off or productive people of a particular social scheme or society and Group N represents the group of less well-off or unproductive people of that same social scheme.

Figure 4.2: **Matrix Showing Cooperation and Noncooperation between Better-off Group S and Less-well off Group N**

	Group S	
	Don't Secede	Secede
Don't Secede (Cooperate)	1000, 1600	500, 2000
Group N		
Secede (Don't cooperate)	500, 2000	500, 2000

If Group S does not secede (i.e. cooperates with Group N), Group N gains and Group S losses, and if Group S does secede (i.e. does not cooperate with Group N), Group N losses and Group S gains. In the classic or traditional form of the PD, the choice is between cooperation and noncooperation, the optimal choice being cooperation and the equilibrium choice being noncooperation. If the numbers represent utilities, then as the matrix shows, Group N prefers non-secession to secession since non-secession provides it greater utilities (1000 utilities compared to 500 utilities when group S does not cooperate). The contrary holds for Group S, which prefers secession to non-secession because secession provides it higher utilities (2000 utilities compared to 1600 utilities when it supports Group N). Thus, the 'optimal-dominant choice' for Group S is noncooperation and the 'optimal-dominant choice' for Group N is cooperation. Keep in mind though that this is not a standard representation of the classic form of the PD. The situation of Groups S

and N is sketched in the form of the PD here just to bring out the sort of reasoning in terms of utilities or benefits that is going on.

Note that in the matrix the entries are utilities, but utilities of a certain sort, namely, expected utility. Both Groups are motivated to act by these utilities. On the one hand, members of Group N choose the cooperative act because it offers them greater EU. On the other hand, members of Group S choose the secession act because it maximizes their EU. In this account or representation, there is no place for non-EU moral reasons. Reasons that go beyond EU, such as those that speak to acts that produce the outcomes, say, the meaning of secession and cooperative acts, or their value, or productive members' considered preferences or aversions for the acts of secession and cooperation.

To illustrate the point about non-EU moral reasons, take the case of someone (let us call him Jonas) contemplating whether he should keep or return the wallet he found in the library. Jonas is Copp's prototype reserved maximizer, but unlike Copp's description, I have added more information to make Jonas' situation more striking. If Jonas keeps the wallet no one will find out. He is absolutely certain of this.[266] He is in a jackpot situation because there is enough money in the wallet. He knows he has much need for extra money. His wedding is right around the corner. He has bought most of the things that are needed, but they are a few important items, say, the wedding ring and the wedding cake that are still pending. He hopes to buy these if he gets the vacation bonus from his employer, but he is not sure whether he would get it before the wedding.

[266] Shall we say that Jonas is in this particular instance in possession of the Ring of Gyges?

Should Jonas keep the wallet or return it to lost and found? Would it be rational for him to return it or keep it? Suppose he chooses to keep the wallet what would Mb(CM)A say? Because Mb(CM)A is an EU-focused account its evaluation of the rationality of Jonas' choice to keep the wallet is typically based on what he gains from the possible outcomes of the acts that are available to him. If Jonas returns the wallet he losses the money and if he keeps the wallet he doesn't lose the money. If Jonas gains by keeping the wallet, then Mb(CM)A will consider his choice rational—the act offers him greater EU. Conversely, since he losses by returning the wallet, Mb(CM)A would consider his choice irrational—the act offers him fewer EU. For Mb(CM)A, when EU is stacked too high against returning the wallet to lost and found, it is not rational to return it.

Now, it is obvious that something significant is missing from this account. It is a single-tracked silver bullet explanation of the situation and of Jonas's reasons for acting. The account does not take into account Jonas' considered preference or aversion for the acts that are available to him. What would it mean for an account to factor in Jonas' considered preference or aversion for the acts of 'keeping the wallet' or 'returning it to lost and found'? We would expect the account to describe these acts in ways that give different weights to them. The weight-bestowing account or description takes into account Jonas' considered preference or aversion for the acts and it does not tell us straight away whether it is rational or not for him to keep the wallet.

There is a fundamental reason as to why a silver-bullet explanation of Jonas' situation is misleading. Jonas' situation like the problem of secession and most

Prisoner Dilemmas or choice situations require silver-bullet explanations. The situations require silver-bullet explanations because they present agents with options, strategies, or actions that have to be bestowed various weights. In these situations the agent does not just ask, what is in it for me, i.e. what options, strategies, or actions maximize expected utility? In addition to asking this the agent asks, what options, strategies, or actions symbolize *x* value? Where *x* is a particular value that the agent holds.

A weight-bestowing account does not claim that it is rational for Jonas to keep the wallet. Neither does it claim that he acted irrationally just in case he chooses to return the wallet. On the contrary, a weight-bestowing account claims that the rationality of Jonas' choice depends on what the various acts symbolize for him, in addition, of course, to their possible outcomes. If the act of returning the wallet expresses, for him, say, the value of honesty, then, according to a weight-bestowing account, it would be rational for him to keep it, factoring in the expected utility of the act. Conversely, if the act of returning the wallet does not represent for him any such value, then it would not be rational for him to keep it, factoring in the expected utility of the act.

Hence, if Jonas returns the wallet to lost and found, he does so not simply because of expected utility, but because of his considered preference for the act of retuning it and his considered aversion for the act of keeping it. And if he keeps the wallet, he does so not merely because of the possible outcomes of the acts, but because of his considered preference for the act of keeping it and his considered

aversion for the act of returning it. This weight-bestowing account is a DV/SU account, the sort that I shall be arguing for in chapter five.

4.2.1 Kavka's Three Solutions to the Scope of the Contact Problem and the Problem of Secession

In this section, I shall examine Kavka's three solutions to the scope of the contract problem and the problem of secession. Of the three solutions he discusses, one directly appeals to expected utility; the other two appeal to general affectivity and practical considerations. These solutions explain "why the endowed, the handicapped, and so on would probably not prefer to secede from the negotiations."[267]

The first solution restricts the formation of coalitions during negotiations. Let us call this solution the 'rationality of coalition formation solution.' This restriction, according to Kavka, "does not absolutely preclude secession of a subgroup intending to set up its own commonwealth"; however, "it does inhibit the formation of such groups during the negotiating process by forbidding the coalitions from which secession groups would be most likely to arise."[268] By interfering with "possible agreement by a subgroup on an alternative social contract prior to withdrawal from the full-group negotiations," the restriction "introduces a substantial element of risk"[269] into negotiations and secession; the risk being that would-be secessionists would be left without any satisfactory agreement and a social contract.

[267] Kavka, p.241.
[268] Ibid, 241.
[269] Ibid, p.241.

Note that the restriction on the formation of coalitions limits the choice of contractors during the negotiation stage to *one partition*, i.e. a full-group commonwealth or social contract. Moreover, it assumes that this partition provides them greatest gains for contractors. Given that partition *p* offers contactor X greater benefits over alternative partition or social contract X would have a rational motivation to accept it. If X attempts to secede from *p* by forming coalition with others X runs the risk of being excluded *simpliciter* from the bargaining process and from full-group commonwealth. Kavka writes:

> The individual who secedes sacrifices his hopes of inclusion in a full-group commonwealth, without knowing precisely who will join him in secession and whether and on what terms agreement might be reached among those who secede.[270]

Given the risk of exclusion from full-group commonwealth, rational prudence would prohibit any attempt at coalition formation during the negotiation process. Since the high costs of coalition formation would sufficiently deter the endowed or better-off contractors, say, members of Group S from negotiating with other better-off contractors, the contract ends up being one of full-group commonwealth in which agreements are extended to everyone.

A number of things are wrong with Kavka's 'rationality of coalition formation solution.' Firstly, in limiting the choice of contractors to one partition it wrongly assumes that such a partition would be *acceptable* to every contractor. But as Gauthier states, and rightly in my view, "there may be many, mutually exclusive

[270] Ibid, p.241.

partitions, each of which would seem to some persons to be the basis for determining the scope of social contracts."[271] If we suppose that all the possible partitions satisfy the condition of rational agreement, according to which it is not rational for persons to extend their agreement more widely than benefits them, then choosing among which of the partition should form the basis of determining the scope of social contracts would be problematic.[272] Furthermore, since there may not be agreement among contractors regarding which partition ought to determine the scope of social contracts, the claim that rational prudence would prohibit any attempt at coalition formation during the negotiation process seems moot.

Secondly, for better-off contractors to be effectively deterred from seceding or from forming coalition with other better-off contractors, they must be deprived of full knowledge of their characteristics, abilities and circumstances. The 'rationality of coalition formation solution,' Kavka says, does exactly this. He writes:

> In developing Hobbesian theory, then, we shall follow a middle ground between allowing the contracting parties full knowledge of their personal situations and Rawls' strategy of ruling out all such knowledge. In particular, our basic assumption shall be that the parties know their personal characteristics but not their social positions.[273]

By full knowledge of one's personal characteristics Kavka means knowledge that one is, for example, an intelligent success-oriented person. And by full knowledge of one's social position he means knowledge that one is, for example, an "upper-

[271] Gauthier, The Social Contract: Individual Decision or Collective Bargain?" p.63.
[272] Gauthier even assumes that in order to "decide among possible partitions," a theory of rational coalition formation is required, and "whether such a theory is even possible raises issues" that he says he cannot begin to consider, Ibid, p.63.
[273] Ibid, pp.193-194

middle-class attorney earning a high income and having substantial and political influence in her community."[274]

It is not clear however, if the partial veil of ignorance that Kavka imposes on the contractors—in order to screen out knowledge of their social positions—is sufficient to deter them from forming coalitions with others. Since Kavka allows them knowledge of their personal characteristics, it is most likely that contractors will make inference from that to their social positions. In which case a person may believe that a certain personal characteristic, say, an intelligent success-oriented attribute may causally contribute to him or her being a successful person or doing well at whatever career he or she chooses. In other words, an individual who knows that he or she is intelligent or has an intelligent success-oriented attribute would most likely believe that there is a high chance that he or she will do well, career-wise, in whatever society she ends up in, unless of course she ends up in a society that does not reward intelligence, which is highly improbable. Hence, because they know that they possess certain intelligent success-oriented attributes, contractors will move to form a coalition among themselves with the expectation that their intelligent success-oriented attributes will causally contribute to their being successful.

In any case, since Kavka's strategy which deprives contractors knowledge of their social positions runs counter to the view that full knowledge of contractors' characteristics, abilities and circumstances is necessary for the success of rational bargaining and for the motivational efficacy of acting on the principles that are chosen we reject it. Since contractors may prefer a social contract other than the one

[274] Ibid, p.194.

that the partial veil of ignorance has imposed on them getting them to willingly comply with its terms would be problematic. Once we assume that contractors are not hidden behind any veil of ignorance, partial or full, Kavka's argument for the 'rationality of coalition formation solution' falls flat. Given that contractors have knowledge of their characteristics, abilities and circumstances excluding them from full-group commonwealth works to their advantage. If the preference of the endowed and better-off contractors is for a society of endowed and better-off members, leaving them out of full-group commonwealth as punishment for their behavior plays right into their hands since they will naturally move to form a group of better-off members.

Kavka's second solution assumes that "people are predominantly, rather than purely, egoistic."[275] There are three claims embodied in this solution—claims that Kavka did not really argue for. Spelled out fully the claims are as follows:

(1) The non-endowed, handicapped and better-off have well endowed, nonhandicapped and less well-off relatives and friends who will refrain from any "secession movement aimed at leaving [them, i.e. the less well-off] to their own device."

(2) Since the handicapped are generally both needy, harmless and unthreatening, they are likely targets of whatever general altruism or sympathy that is possessed by the parties.

[275] Ibid, p.242.

(3) "The knowledge that some of their children or grandchildren could be handicapped, the parties could be expected to prefer to help support the handicapped."[276]

All three claims conflict with the moral demand of moral contractarianism. They conflict with Mb(CM)A because they appeal to fixed moral affectivity and not to the affective capacity for (rational) morality. Claims 1 and 2 are rejected because they are incompatible with Mb(CM)A's idea of an essentially just society. Recall that the morality of Mb(CM)A is the morality of an essentially just society. An essentially just society—which is chosen from an Archimedean standpoint—encourages the free development of the affective capacity for (rational) morality. But this is exactly what claims 1 and 2 deny. They appeal to fixed moral affectivity, that is they claim that because endowed and better-off members would necessarily have sympathy towards the handicapped and unendowed they "would not join a secession movement aimed at leaving them to their own devices." But if morality speaks only to the reason of rational agents, namely, to agents who have freely developed affectivity, then better-off members would consider it rational not to exclude from cooperation less well-off friends and relatives only if (i) they have unendowed and less well-off friends and relatives, and (ii) they consider it sufficiently in their own interests not to exclude less well-off friends and relatives from cooperation.

The third claim is a variant of Rawls' lexical difference principle, where the least advantaged member of society provides the standpoint for the choice of the

[276] Ibid, p.242.

principles of justice, except in this case the contractors consider not their own interests, but those of their children and grandchildren. I argued in chapter two that in general, if we accept all the assumptions in JaF, then not only will Rawls' two principles of justice be chosen, but his version of contractarianism would constitute a persuasive and fruitful approach to distributive social justice. Because JaF insists on principles that provide a safety net for contractors, we might say it supports the less well-off members, i.e. citizens of the North through a social scheme that redistributes economic resources and the bases of self-respect. The social scheme that Rawls' principles of justice structures guarantee that everyone is respected as equals and not as superiors and subordinates.

The point about contractors in JaF accommodating themselves to various social schemes has been made forcefully by Braybrooke. Unlike Gauthier's agents, Rawls' agents, according to Braybrooke, are in a better position to accommodate variations and changes in capacities and resources since they have an "incentive to hold open the possibilities of following life plans that none of them happen to favor."[277] Furthermore, since Rawls, and not Gauthier uses the standpoint of the least advantaged member as the standpoint for the choice of the principles of justice, he is able to argue for the handicapped, the elderly, the unendowed and less well-off. But Gauthier accepts none of Rawls' assumptions and hence rejects JaF as a compelling theory of distributive social justice. In arguing against JaF, particularly, the standpoint of the least advantaged member of society that underlines it, Gauthier says:

[277] David Braybrooke, "Gauthier's Foundations for Ethics under the test of Application," in *Contractarianism and Rational Choice*, p.59.

> But the argument for the maximin principle is also open to a fundamental objection. …the argument…assumed that our randomly selected individual would identify with the standpoint of the least-advantaged person, in choosing a principle for cooperative action. But his identification is purely arbitrary. A person who finds that in society he is in a more advantaged position will not consider that the maximin principle was rationally selected, simply because it maximized the expected utility of someone else – someone other than the person who he turned out to be. Each party to the contract will insist that, although the decision in the state of nature must be made in ignorance of who he is, it must be reasonable for him, no matter who he turns out to be. In other words, it must be reasonable from every standpoint, and not just that of the least advantaged person.[278]

Beyond the fact that it requires that endowed and well-off members allow others to freeride on their back the third claim can hardly be defended as reasonable for those who may not have children or grandchildren or those who do not think that the possibility of their children or grandchildren becoming handicapped and unproductive is a sufficient motivation for them to support the handicapped and unproductive.

Kavka's third solution, which I shall call the 'no-distinct-territory view,' argues that any preference to secede would be encumbered and frustrated by the nature of the territory that negotiators occupy. The main thrust of the 'no-distinct-territory view' solution is the claim that it is practically impossible for better-off negotiators to form a state among themselves because they most likely do not live in a separate region. It is a priori unlikely, Kavka says, "that the endowed and

[278] Gauthier, "The Social Contract: Individual Decision or Collective Bargain?" p.54.

unendowed negotiators occupy distinct territories."[279] Since negotiators are most highly interspersed within a single geographical region, "secession and agreement by the endowed alone would still leave them with the problem of dealing with the unendowed in their midst."[280] Given the practical difficulty of forming a state of better-off negotiators, the most likely social contract that would emerge from the bargaining process is one that consists of the unendowed and endowed negotiators. However, the argument that any preference to secede would be encumbered and frustrated by the nature of the territory that negotiators occupy losses its force when we consider situations where negotiators are not intersperse within a single geographical region. Braybrooke discusses one such situation. I turn to him for an explication of this.

4.2.2 Braybrooke, Mb(CM)A, and the Problem of Secession

Braybrooke illustrates the acuteness of the problem of secession with a society that was once united by interdependence of utilities and free affectivity. However, due to the withering away of earnings of inhabitants of one part of the country, they are no longer able to make a net contribution to the production of public goods. What should members of the affluent, productive, and well-off part do in this situation? Braybrooke writes:

> In a society formerly embracing everyone with every other in a relation of mutual advantage (let us say, specifically earning their keep in private goods and making a net contribution to the production of public goods),

[279] Kavka, p.241.
[280] Ibid, p.241.

> developments occur that make a substantial part of the population redundant, so far as contributing anything of value to the remaining part goes. On the contrary, the redundant part, who inhabit the North, drain away goods from the South to support them in idleness. What would the people in the South make of that? The interdependence of utilities that along with free affectivity once united them with their now redundant fellow contractors would no longer have a foundation…. Would it not be rational for them, and in accordance with Gauthier's assumptions to form a coalition and secede from the larger society?[281]

The society described by Braybrooke above is divided by their economic activities. Members of this society are not interspersed within a single geographical region. Let us say the South is industrial and the topography in the North supports agrarian activities. The wave of development, for some reason, has left them redundant and unproductive and as such they can no longer contribute to the cooperative surplus or the production of public goods.

Braybrooke talks of the above society in the abstract; however, there is much it has in common with the Nigerian society. The vastness of Nigeria's resources previously came from the North region. However, this changed considerably about a decade after the country's independence—from British rule—following the discovery of oil resources in the South region; revenue from these resources is used for the benefit of the country, including the North region. So economically, Nigeria depends on the South region. But politically, the North region dominates, primary owing to its slightly higher population. To be sure, there is an imbalance and asymmetry—politically and economically—in Nigeria; but

[281] Braybrooke, "Gauthier's Foundations for Ethics under the test of Application," in *Contractarianism and Rational* Choice, p.65.

most importantly, the society fails the demand of mutual advantage because the North region lives off the productive activities of the South region. The question then becomes 'would it not be rational for the South, and in accordance with Gauthier's assumptions to form a coalition and secede from the larger society?'[282]

Note that the society that Braybrooke describes for which Nigeria provides a concrete and practical example is divided by its economic and productive activities. Furthermore, members of this society are not interspersed within a single geographical region. Thus, contra Kavka's 'no-distinct-territory view' solution there would be no practical difficulty for secession. If the better-off region chooses to secede from the larger society given that it occupies a distinct geographical region it would be unproblematic for it to do so.

Evidently, the society violates the demand of mutual advantage; hence, there is much to be said about the rationality of cooperation. Given that mutual advantage is a necessary condition for cooperation and given as well that the Nigerian society does not meet the condition, it would be rational, according to Mb(CM)A, for citizens of the South not to support citizens of the North. Gauthier's contract theory, "like the morality directly entailed by it," as Braybrooke puts it, "leaves out of account people who are not in a position to contribute to producing any part of the cooperative surplus."[283] Mb(CM)A has delivered its judgment and the verdict is loud and clear: unproductive and less well-off members from the North should be excluded by productive and better-off members from cooperation.[284]

---

[282] As things presently stand in Nigeria they are elements in the South region that desire to separate from the federation and are systematically pushing for that.
[283] Braybrooke, "Social Contract Theory's Fanciest Flight," in *Ethics*, p.756.
[284] Gauthier, "The Social Contract: Individual Decision or Collective Bargain?" p.62.

In various places in MbA and elsewhere, Gauthier argues powerfully that the handicapped and defective, the unborn, the dependent elderly—and we can add the unendowed or simply put those who neither have nor no longer have anything left to contribute to the cooperative surplus—"fall beyond a contractarian morality, a morality tied to mutuality."[285] As far as Gauthier is concerned, the best way to breed freeriders or to violate the thesis of individualism is not to take seriously the demand of mutual advantage. Since members of Group N no longer have anything left to contribute to the cooperative surplus because of their unproductivity it is not rational or in the self-interest of members of Group S to support them. Simply put, it is EU-rational for the endowed and productive part or members to secede from the larger society. Whereas the demand for secession by citizens of the South is similar to the demand of members of Nation S—and on the basis of the requirement of mutual morality they are justified, according to Gauthier, to argue for a contract of national society that protects their interests—the insistence for support by those from the North is similar to the demand of the members of Nation N that the social contract ought to be based on the society of human race. In both cases, it is not EU-rational, according to Gauthier's brand of contractarianism, for better-off members to extend support to less well-off members.

For Braybrooke, the fundamental issue is that since all that matters for Gauthier's agents is the maximization of expected utility they would care less among other things, on how the less endowed and less well-off fare in society. The commitment of Mb(CM)A to the maximization of expected utility, notwithstanding its modification of the conception of rationality of standard rational choice theory

[285] See MbA, p.18 fn.30, and p.268.

means that in the language of cooperation only actions or dispositions that maximize expected utility matters. Gauthier's assumption is that agents have a rational motivation to be disposed to cooperation or to support a scheme of cooperation insofar as cooperation maximizes their expected utility. Since it is not in an agent's self-interest to extend agreement more widely than benefits him or her, it is not rational for that agent to remain in a scheme of cooperation that fails to maximize expect utility.

There is the worry of instability that confronts Gauthier's brand of contractarianism. The worry is that to prohibit agents from extending their agreement more widely than benefits them is to require them to seek cooperation with only a limited productive and better-off few. Moreover, there is the point about Gauthier's assumption opening a 'Pandora Box of secession.' Suppose that members of X-group successfully secedes from *N* society on the ground that the society does not maximize their expected utility. Suppose also that after seceding X society is formed. Suppose finally that after some years development and changes occur in X society that leave some part of that society redundant, unproductive, and some members less well-off. Given that all that matters for Gauthier's agents, for rationality, and for cooperation is the maximization of expected utility, it would be rational for the more productive part to secede in order to form a society among themselves. But suppose that the productive part of X-society successfully secedes. Now, suppose also that after seceding part of the new society becomes unproductive and redundant, shouldn't the better-off agents secede too? And shouldn't the better-

off agents of the newer society secede and the better-off agents of the newest society secede, and so on.

But why, Braybrooke ask, "should we think that all that any of us want to find in morality or justice will be put there by reason alone—by merely rational agents, even rational agents ready on utility-maximizing grounds to put themselves under the constraints required for cooperation?"[286] In directing his question to Gauthier's theory that founds morality on rationality Braybrooke legitimately asks, Can reason give is an adequate morality? And if yes, does it give us enough resources to handle serious problems like that of secession?

It is important to point out that Braybrooke's question raises the issue of insensitivity for rational morality. Now, in general, it seems unsympathetic for an individual to be indifferent to the plight and condition of anyone in need. We may suppose that individuals who have suffered some various forms of economic tragedies and problems need the support of those that are better-off. We may further suppose that for those that are better-off to hold back such support from those in economic need is to demonstrate not only a lack of understanding but to fail where it morally matters.[287]

Such support or affectivity appeals, according to Braybrooke, to natural and apposite moral sentiments. Whereas Braybrooke takes affectivities to arise from some prior moral obligations, Gauthier takes them to arise from the preferences of agents. For Mb(CM)A, affectivities do not arise from prior moral obligations because for it *morals are by agreement*. Moral norms and for that matter

[286] Braybrooke, "Social Contract Theory's Fanciest Flight," in *Ethics*, p.756.
[287] Such support is not limited to the emotional; it includes as well material assistance.

affectivities for Gauthier's theory are not grounded on some prior moral obligations but on a conception of rationality that is expected utility-sensitive.

But don't the past contributions by the less well-off justify supporting them? Shouldn't the past contributions to the cooperative surplus by the less well-off impose some particular obligation on the rest of society? It could be argued that a particular obligation is imposed on the rest of society in virtue of the past productive efforts of the less well-off agents. Isn't this the reasoning behind superannuation and pension benefits?

There is certainly a significant moral difference between supporting a group of unproductive people who have never contributed to the production of public goods and those who had in the past contributed to the cooperative surplus. It would seem that the contributions of the latter impose a prima facie obligation on the rest of society to support them. If the past contributions to the cooperative surplus by less well-off agents impose a particular obligation on the rest of society, then for Group S to exclude Group N from the social contract because they are no longer in a position to contribute to the cooperative surplus treats them not only cruelly, but ignores this particular obligation. Having contributed to the cooperative surplus for many years, as productive members of society it would be morally wrong to use their present unproductiveness as a reason to cut them off from the benefits that cooperation offers.

Rather than argue that the past contributions to the cooperative surplus by the less well-off impose some particular obligation on the rest of society, we might argue that better-off agents have an indirect duty to support less well-off agents

insofar as the duty engenders the development of character-traits that contribute to human flourishing. This argument is similar to Kant's argument that humans have indirect duties to treat non-human in such a way that is morally good for humanity. Kant's moral theory categorizes duties into two: those that are perfect and those that are imperfect. The former are violated when the maxims that an individual acts on lead to a practical or logical contradiction, while the latter is violated just in case it leads to a conflict of rational willing. Both duties apply to rational persons as lawmaking members of the kingdom of ends.

Although in Kant's moral ontology, perfect or imperfect duties to non-human animals are ruled out since these are duties that are applicable only to lawmaking members of the kingdom of ends, he does believe that we have indirect duties to treat non-human animals in such a way that is morally good for humanity. He writes:

> But so far as animals are concerned, we have no direct duties. Animals are not self-conscious and are there merely as a means to an end. That end is man. Our duties towards animas animals are merely indirect duties towards humanity.[288]

These indirect duties towards humanity derive from the duty to strengthen the feelings of compassion. It is for this reason that it would be unreasonable to go around kicking or mistreating dogs, cats or other non-human animals. Using the same reasoning it could be argued that better-off agents have an indirect duty to less well-off agents; they ought to support less well-off agents insofar as that

[288] Kant, "We Have No Duties to Animals," from *Lectures on Ethics*, trans. Louis Infield, London, Methuen Press, p. 239.

strengthens in them certain character traits, say, adaptability, compassion, perseverance and self-control that are essential for human flourishing. Since these traits are necessary for a life well-lived rational agents would seek to develop them as possible as they could.

From the foregoing two comments are in order. First, since for the most part, social institutions and practices have relatively held up against instability and issues of secession, we might suppose that individuals do not take their reasons for acting solely from expected utility. Stated differently, the relative stability of society may in fact be a whiff of support for the view that members of society have been socialized to include among their reasons for acting considerations that extend beyond expected utility. Second, if EU-reasons and non EU-reasons factor into the decision of people regarding what acts they choose, then the claim that Mb(CM)A's conception of rationality is narrow and misleading, and in need of modification seems right.

Note that in raising the test of application for Mb(CM)A, Braybrooke puts forth two theses: a negative thesis and a positive thesis. The negative thesis claims that (a) Mb(CM)A cannot resolve the problem of secession, and (b) any social contract theory cannot resolve the problem of secession. The positive thesis claims that only a theory of moral or receptive sentiments resolves the problem of secession.

Braybrooke's argument for the two parts of the negative thesis is that Mb(CM)A or any social contract theory of reasons fails in the test of application because it wrongly assumes "that all that any of us want to find in morality or

justice will be put there by reason alone."[289] The positive thesis follows from this argument. According to this thesis, reason alone does not give us all that we want to find in morality or justice; reason might give us some, but it is moral sentiments that give us what we want to find in morality or justice. Consequently, only a theory of moral sentiments resolves the problem of secession.

Braybrooke states what seem to be both the negative and positive theses in the following passage:

> Gauthier's theory may in fact run against humane feeling as much as it runs with it…. In the accumulation of exceptions and in other processes for changing social rules, the main action for moral progress lies elsewhere, in activity prompted more by sentiment than by theory…. Contract theory may teach us how to guard against licensing sloppy results; but contract theory cannot do the moral work for which, in the foundations and applications of ethics, sentiments has been appointed.[290]

Whereas Gauthier's theory in particular and social contract theory in general fails where it matters most (in the foundations and applications of ethics), a theory of receptive sentiments, Braybrooke argues, succeeds.

As Braybrooke sees it, a theory of apposite sentiments or some rich account of affective moralities and not a contract theory is able to dissolve issues of secession and moral progress. Because Mb(CM)A is a morality of reason rather than of sentiments, it cannot do the moral work for which, in the foundations and applications of ethics, sentiments has been appointed. For all we want, we may hail

---

[289] Ibid, p.756
[290] Braybrooke, "Gauthier's Foundations for Ethics under the test of Application," in *Contractarianism and Rational* Choice, p.70.

Mb(CM)A for its rigorous, systematic and strikingly sophisticated approach to moral and political theorizing. If we want, we may appreciate it for its high-flying unpacking of rationality. And if we desire, we may salute Gauthier for "the technical virtuosity that he displays [in fashioning out a rigorous and systematic rational morality], "and the threefold philosophical triumph that he achieves" in doing this.[291] But we cannot, Braybrooke says, rely on it to provide us the appropriate apparatus to resolve issues of secession.

For sure, Mb(CM)A is not grounded on affectivity or sentiments and to this extent, Braybrooke is right to claim that it might be problematic as an account of social justice when it comes to issues of secession. We agree with negative thesis (a). Could it be that overall we have oversimplified or even misstated Mb(CM)A's take on rationality explained by expected utility and its application to issues of secession? Perhaps, the mistake is in thinking that actors in Mb(CM)A have a small threshold level, that is they necessarily 'walk away' in situations were there is a threat to their utility profile. An individual who, for example, is involved in a marriage union, and whose utility profile has been threatened, might be more congenial to many more rounds of bargains and negotiations with his or her partner than he or she would be, were he or she deadlocked with his or her soccer club or employer. If this is right, then we should expect that contractors would hold out much longer in some situations (of secession) that threaten their utility profile.

What will it mean for the problem of secession if we accept the above line of reasoning? Not much I should say. At best what the reasoning shows, I think, is that in some situations, individuals are more willing to negotiate and renegotiate

[291] Braybrooke, "Social Contract Theory's Fanciest Flight," in *Ethics*, p.751.

when cooperation or their utility profile is threatened. And even in these situations the motivation for rational actors in Mb(CM)A would seem to be EU-driven. We might interpret such motivation in terms of hope, namely, the expectation that holding out much longer might lead to a better and favorable situation. If the reason why some hold out much longer in, say, a marriage relationship that is threatened is that they expect to be able to fix things and if this reason is EU-driven, then it is reasonable to claim that if this hope of fixing things gets extinguished, they would no longer have any motivation to hold out. Thus, non-cooperation would still be overwhelmingly attractive in these kinds of situations given that all that matters for actors in Mb(CM)A is the maximization of EU. Again, when EU is stacked too high against cooperation, whether in situations where secession looms large or in the PD, or other choice situations, it is not rational, according to Mb(CM)A, to cooperate.

I thus agree with negative thesis (a), but not negative thesis (b). I believe and I shall argue that the Mb(DV)A account I defend, which is a modified moral contractarian account of reason is able to dissolve the problem of secession. It is true that both accounts (Mb(CM)A and Mb(DV)A) are theories of reasons. However, the reason for Mb(CM)A's breakdown in the test of application is that it appeals to EU-reasons. On an Mb(DV)A account that I shall be defending in chapter five, the reasons that are appealed to are not limited to EU, they include as well reasons about value, namely, reasons that appeal to the meaning of the acts that produce the outcomes.

On a side note, one wonders why Braybrooke thinks that contract theory cannot do the moral work for which, in the foundations and applications of ethics,

sentiments has been appointed. JaF is a social contract theory; it is a theory of moral reasons and not of moral sentiments. And as I have shown at the outset of this chapter, JaF fares well when applied to the problem of secession—of course, once we accept its various assumptions—because it subscribes to principles of justice that structure the basic institutions of society in such a way that the unendowed, unproductive, and less well-off are provided a safety net. If JaF's egalitarianism and strict redistributionist outlook prevents a situation where the endowed, productive, and better-off members secede from a scheme of cooperation involving them and the unendowed, and unproductive, then the second part of Braybrooke's negative thesis seems misleading if not patently false.

For all we know, Braybrooke may be right that a theory that is grounded on moral sentiments is a more plausible approach to the problem of secession. However, given that the affectivities he defends conflicts with Mb(CM)A's idea of an essentially just society, I do not accept his approach to the problem of secession. But can Braybrooke's positive thesis be defended? In particular, can it be shown that a theory of moral sentiments is able to resolve the problem of secession? How would a theory grounded on affective morality fare in the test of application? I examine this is in the next section.

To summarize, the problem of secession is a problem that arises within the life of the contract. It is a problem concerning what to do with previously productive members of society who for some reason have become unproductive. Mb(CM)A's conception of rationality, which takes expected utility to be basic informs its solution in the test of application. In agreeing with the first part of

Braybrooke's negative thesis I claimed that Mb(CM)A's characterization of rationality is inadequate and its solution to the problem of secession is misleading. I examined Kavka's three solutions to the scope of the contract problem and the problem of secession, and I rejected all. One of the reasons for rejecting the 'rationality of coalition formation' solution is that it encumbers the bargaining process by imposing a partial veil of ignorance on the parties. The second solution was rejected because it is impractical when applied to Braybrooke's society for which Nigerian provides a practical and concrete example. And I rejected the 'no-distinct-territory view' solution on the ground that it conflicts with the foundations of Mb(CM)A. Specifically, it conflicts with Mb(CM)A's view that the ground of obligations are not fixed affectivity. Affectivity does not provide the reasons and ground for acting for agents for whom an essentially just society, filtered by reason and anchored on free affectivity is well suited.

4.3 Affective Morality and the Problem of Secession

Affective moralities are distinguishable from rational moralities primarily in the way they specify the motivation for actions. Affective moralities identify the reason for acting along the dimension of sentiments or feelings. In contrast, rational moralities identify the reason for acting along some rational dimension, which is specified in terms of either human reason or rational self-interest. In discussing affective moralities, I shall be limiting myself to Hume's moral theory or theory of moral sentiments. In focusing on Hume's theory of morality my objective is informed by Braybrooke's positive thesis: that a social contract theory cannot do

the moral work for which, in the foundations and applications of ethics (the problem of secession), a theory of moral sentiments has been appointed.

4.3.1 Hume as a Contractarian?

Some have interpreted Hume along contractarian dimensions. Gauthier, for example, claims that notwithstanding Hume's stated disavowal of contractarian thought, he is a contractarian at bottom. His argument is that although Hume's moral theory is founded on moral sentiments, his theory of property and justice and the theory of government and obedience are contractarian in their rationale, since mutual advantage or common interest is their only condition.[292] Let us call Hume's theory of moral sentiments MS, his theory of property and justice PJ, and his theory of government and obedience GO, Gauthier's argument is that whereas MS identifies primarily with receptive sentiments, PJ and GO possess a strong contractarian component because they identify with mutual advantage.

But how can this be since Hume treats justice as a moral virtue? If it is the case that justice is a moral virtue we would expect that MS would be connected with PJ and GO. Gauthier's argument is that MS might be connected with PJ and GO, but 'connection is not identification.'[293] Now to be sure, if Gauthier can pull off the argument that Hume is some sort of a contractarian and his view of property, justice, and society is colored in contractarian terms of the kind that is grounded on mutual advantage, then it might be possible for him to argue that Hume's theory of moral sentiment has a focus different from those of PJ and GO. In which case,

[292] See Gauthier, "David Hume, Contractarian" in *Moral Dealing: Contract, Ethics and* Reason, pp.45-75.
[293] Ibid, p.45.

moral sentiments apart, what matters for Hume is that in the progression of society our interactions with others be just and fair. The justice and fairness of such interactions being determined by the extent they are mutually beneficial.

Since Gauthier thinks that identification of PJ and GO with contractarian thought does not undermine the non-contractarian character of MS, it would be fruitful to examine the sense in which PJ and GO, or Hume can be interpreted along a contractarian dimension. Gauthier begins with Hume's view of public utility, which as he rightly notes, is the sole origin of justice and common interests underlying the establishment of governments. Gauthier interprets both public utility and common interests as mutual advantage and this, he argues, is a cardinal condition or feature of any contractarian enterprise.

To get a sense of the sort of interpretation and decoupling of Hume's view that Gauthier is suggesting we need to distinguish five senses of contractarianism, four of which Gauthier discuses.[294] First, *original contractarianism*, the view that the origin of society, property and agreement is to be found in a contractual convention among humans. Second, *consent* or *explicit contractarianism*—the theory that defends property and political institutions by appealing to actual agreement among members of a political society. Third, *tacit contractarianism*, which claims that the acceptance of the benefits provided by systems of property and political institutions confers legitimacy on them since the acceptance of these benefits is a form of tacit consent. Fourth, *hypothetical* or *analytical contractarianism*, according to which the practices and institutions of society are justified if they would be the object of agreement among rational persons in a

[294] Ibid, pp.52-54.

suitable choice situation. Fifth, *virtual contractarianism* (defended by Harman),[295] which justifies the terms of the contract by a moral system's rights and duties. The contract, on this view, is interpreted as if it lies behind a moral system of rights and duties.

Hume's anti-contractarian avowal in MS, Gauthier asserts, targets the first three senses of contractarianism and not the fourth (and possibly not the fifth sense). Hume, Gauthier says, disavows contractarianism only as it is interpreted along the first three senses. His contractarian disavowal in MS and his views regarding consent and interests are not incompatible with the fourth sense of contractarianism (hypothetical contractarianism).[296] Since PJ and GO subscribes to mutual advantage as its only condition for justification—a condition that is cardinal to hypothetical contractarianism—Hume's moral and political inquiries, Gauthier argues, lean towards a contractarian interpretation. How plausible is this?

Dario Castiglione has argued that Gauthier's contractarian interpretation of Hume seemingly confuses part of Hume's theory with the whole and wonders if such interpretation is not a speculative exercise, lacking perhaps any deep anthropological foundation.[297] Morality, for Hume he argues, "is a mixture of natural and artificial virtues"[298] and since neither can completely prevail, any attempt to decouple them would be to fail to understand Hume holistically. A world, according to him, in Hume's view, "where everyone is motivated only by

[295] Gauthier did not discuss this form of contractarianism.
[296] See Ibid, pp.56, 57.
[297] See Dario Castiglione "History, Reason and Experience: Hume's Arguments against Contract Theories," in *The Social Contract from Hobbes to Rawls*, David Boucher and Paul Kelly (eds.), London, Routledge, 1994, pp.108-111.
[298] Ibid, p.109

artificial virtues is not a *moral* world." On the other hand, "a world dominated only by natural virtues is not a *viable* world."[299]

In general, Gauthier may agree with the problem confronting a view that decouples MS from PJ and GO. But he might insist that given that Hume in PJ and GO places a great deal of emphasis on mutual advantage or common interest in curbing individual interest—a view that is largely contractarian—one appears justify to interpret him as a contractarian, notwithstanding his views in MS. In particular, since mutual advantage, for Hume, plays a key role in circumstances of justice or in situations of market failure and informs broadly the nature and extent of our interactions with others when it comes to the maintenance of the conventions of property it seems right to interpret him as a contractarian of some sort.

In discussing Hume's account of moral obligation, Gauthier reminds us that obligation for Hume "arises from a coincidence between an object of our moral sentiments and an object of our reflective self-interest."[300] Gauthier's point about the correspondence between objects of moral sentiments and reflective self-interest is that, for Hume, we are moved to adhere to conventions because of the combination of our moral appropriation and our rational self-interest. If a convention is generally useful, it receives our moral appropriation. As long as we have an 'interest in general conformity to such convention, this interest combines with our moral approbation to give rise to a sufficient moral ground for our adherence to the convention, provided others adhere as well.'[301] So take the example of the artificial virtue, justice and the conventions of property. Justice is

[299] Ibid, p.109.
[300] Gauthier, "David Hume, Contractarian," p. 67.
[301] Ibid, p.67.

understood as "the virtue necessary to the maintenance of the conventions of property." Insofar as the convention of "property is generally useful, justice," according to Gauthier, "receives everyone's moral approbation." But insofar we have 'an interest in maintaining the system of property,[302] our interest combines with our approbation to make justice morally obligatory for us.'[303] Thus, Gauthier says:

> Hume's account of our obligation to be just, to conform to the conventions of property, is thus not purely contractarian, insofar as it reflects his theory of moral sentiments. But insofar as it also reflects his theory of property, it has a strong contractarian component.[304]

4.3.2 Virtues in Hume's Moral and Political Inquiries

I now want to discuss the role of virtues in Hume's moral and political inquiries as they relate to issues of affective moralities. I believe this would help us to understand whether Hume is a contractarian at bottom, or just a 'superficial contractarian,' or a through and through contractarian. Furthermore and more importantly, the discussion would put us in a better position to examine Braybrooke's claim that while a theory that is grounded on moral sentiments is adequately equipped to respond to issues of secession and moral progress, a contractarian theory grounded on reasons is not able to.

---

[302] The lack of such interest in maintaining the system of property or in generally adhering to conventions perhaps marks the difference between the 'Sensible Knave' and the person of justice or moral feeling in situations of circumstances of justice. For the latter, obligation arises when an object of our moral sentiments coincides with an object of our reflective self-interest, whereas for the former, obligation arises only when reflective self-interest is present. See Hume's *Enquiry Concerning the Principles of Morals*, sec. IX, pt. II.

[303] Ibid, p.68.

[304] Gauthier, "David Hume, Contractarian," p.68.

Every action of a moral agent, according to Hume, is motivated by character traits, which are either virtuous or vicious. Virtues are either natural (or instinctive), or acquired (or artificial). The former include charity, generosity and benevolence, gratitude and friendship, while the latter include justice, allegiance, chastity, promise keeping, modesty, and good manners. Artificial virtues, he says are those virtues "that produce pleasure and approbation by means of an artifice or contrivance, which arises from the circumstances and necessities of mankind."[305] Hume takes natural virtues to be the 'more refined and completed forms of those sentiments that are associated with people who belonged to no society but cooperated only within small familial groups.' Natural virtues are natural in the sense that they are not artificially instilled in the agent, who posses them.

Both natural and artificial virtues combine to produce actions and Hume thinks it is not always easy to determine whether a person's motivating character trait is natural or artificial, or perhaps even both.[306] If I drive 25 kilometers just to pick you up from work, it may not be easy to tell whether I did this because I am motivated by artificial character traits, i.e. because I believe it is the just thing to do, or because I promised I would pick you up after work, or because I am motivated by natural character traits, i.e. because you are my friend, or because picking you up is my way of showing appreciation to you for what you have done for me, or because of some or all of the above. We can however, judge if the act is virtuous or

[305] Hume, *A Treatise of Human Nature*, Bk. III, Pt. I, sec. II, L. A. Selby-Bigge (ed.), Oxford, Clarendon Press, 1978, Bk. 3, Pt.2, Sec.1, p.477.

[306] The relationship of both natural and artificial virtues to actions for Hume may be likened to the relationship that Kant assumes holds between intentions and actions in the sense that although intentions move us to act, we may be unable to determine specifically what intentions are responsible for actions or precisely put, what an agent's intentions are when he or she chooses certain actions..

vicious in virtue of our feelings toward it, or the agreeableness of that act from the perspective of the spectator and receiver.

A sense of virtue, Hume says, is nothing other than to "*feel* a satisfaction of a particular kind from the contemplation of the character."[307] This feeling, which could be pleasure or pain constitutes our praise or admiration. When it is praise, the act is virtuous and when it is pain then it is vicious. Virtues are approved because of their utility or usefulness to the spectator and useful actions are approved because of the spectator's ability to sympathize. However, morality or moral approval of actions is not a judgment of reason but an emotional response.

That Hume locates morality in the domain of sentiments in general and in sympathy in particular is not surprising since his moral outlook is framed in opposition to rationalistic moral outlooks. Morality, for Hume, cannot be construed in rationalistic terms. This is because reason is not suited for morality since the rules of morality are not rules of reason. Morals, he says "excite passions, and produce or prevent actions."[308] Reason he boldly proclaims, "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."[309] Whereas reason is for the discovery of truth and falsehood, passion is suited for morality. Truth and falsehood consist in the agreement or disagreement either between the real relations of ideas, or between real existence and matter of fact. In contrast, morality consists in the agreeableness of actions to sentiments. In this sense, actions cannot be said to be reasonable or unreasonable; rather they can be said to laudable or blamable.

---

[307] Hume, *Treatise*, 3.1.2, p.471.
[308] Ibid, 3.1.1, p.457.
[309] Ibid, 2.3.3, p.415.

All virtues, natural or artificial, Hume declares, have to be approved by the spectator. The receiver's agreeable feeling towards an act in virtue of its usefulness to her and the spectator's sympathetic experience of agreeable feelings of the act are confirmation of the virtuousness of the act. This needs further explanation. For although a spectator is naturally disposed to sympathetically approve of any course of action that is useful or agreeable to the receiver, there is a difference between the motivation underlying her approval of natural and artificial virtues. On the one hand, acts arising from natural virtues may bring about her sympathetic pleasure and approval. On the other hand, acts arising from artificial virtues bring about her sympathetic virtue insofar as they reflect a general scheme of advantageous and beneficial circumstances. We are now in a position to say something about interpreting Hume as a contractarian.

To the extent that actions, practices, and virtues like justice, promise keeping, benevolence, and friendship are approved because they are indexed to beneficial outcomes that agents receive, Gauthier seems right to lean towards a hypothetical contractarian interpretation of Hume's moral and political inquiries. Justice, promise keeping, benevolence, and friendship are mutually advantageous and may be generally useful and so receive moral approbation. For if usefulness or utility to the agent is crucial to her moral approbation of these actions, then we may suppose that rational persons would only consider the actions, practices, and institutions of society justified if are mutually advantageous or if they advance the interest of every rational person. And to the extent they are mutually advantageous they would be accepted by every rational person.

However, considering that for Hume, moral motivation and actions are a mixture of natural and artificial virtues, we would expect that natural and artificial virtues like charity, benevolence, generosity, justice, modesty and good manners would play an important role in what an individual decides to do in conditions of extreme scarcity or in circumstances of justice. The individual motivated by a combination of natural and artificial virtues is not primarily motivated by EU-reasons or by mutual benefits, but by these virtues: artificial and natural. Specifically, the individual, in conditions of extreme scarcity or in circumstances of justice is moved by the agreeableness of the situation to her sentiments moderated by the virtues. In which case, Castiglione seems right to suggest that Gauthier is mistaken to decouple the role of both natural and artificial virtues in social interactions, and to disconnect MS from PJ and GO. If we accept the view that both natural and artificial virtues combine to produce actions, then what motivates a better-off agent to cooperate with those who are less well-off cannot be cashed out strictly in contractarian terms or in the language of reasons or mutual advantage. In any, case, even if we suppose that Gauthier's decoupling is a plausible interpretation of Hume's moral and political views, it is not quite clear how useful such interpretation would be for Mb(CM)A when it comes to the problem of secession.

Hume, as I remarked a moment ago, takes the moral approbation of virtues to be dependent on utility, where utility is explained by the capacity to sympathize. If the moral approbation of utility is indexed to the ability of agents to sympathize, then in considering the overall morality of an action, we must factor in the agent's

ability to sympathize. What would it mean to recognize an agent's ability to sympathize in moral situations and in considering the utility of actions? To understand this we need to pay attention to Hume's comments that morality consists in the sympathetic transmissions of feelings or affections from one person to another.

The sentiments of moral approval or disapproval, Hume says, are caused by the operations of sympathy. Sympathy is an emotional or psychological mechanism that enables one person to receive by communication the sentiments of another.[310] Hume states:

> Now the pleasure of a stranger, for whom we have no friendship, pleases us only by sympathy.... sympathy is a very powerful principle in human nature, *that* it has a great influence on our tastes of beauty, and *that* it produces our sentiments of morals in all the artificial [and natural] virtues.[311]

The operation of sympathy begins from the observation of the effects of another person's affection and its outward circumstances or expressions. This observation conveys the idea of passion into our mind. Regardless of their differences, all human beings, Hume argues, are generally similar in their body and in their possession of passions. There are also similar when it comes to possessing vivid impressions of themselves. The vivacity of impressions and our awareness of it along with the principle of resemblance enable us to do more than extrapolate from our situations to those of others. It enables us to receive by communication the sentiments of others. So, when we see others or strangers exhibit particular traits or

[310] See Ibid, 2.1.9, pp.316-324; 2.2.4-7, pp.351-571; 3.3.1, pp.576-591.
[311] Ibid, 3.3.1, pp.576-578.

in situations similar to us, the vivacity of this perception and impressions is transferred to us by resemblance and we come to actually experience passion, and share in their affections.

The traits or situations could be one in which they benefit, or of enjoyment, or it could be one of disadvantage, or harm. If it is the former, we feel enjoyment as well because it is beneficial or agreeable to them (the receivers) and if the latter we feel uneasy because it is harmful or disagreeable to them. We thus feel pleasure because they are pleased, in the same manner we feel pleasure when we experience an aesthetic enjoyment of a well designed object that is not ours. This pleasure or feeling of enjoyment is caused by sympathy. The same, Hume says, can be said regarding laws, social practices, and institutions that we approve of at all times because of their tendency to benefit the whole society of that time or place even where our interests is not at stake. Our approval of this is explained in virtue of our sympathy, the feelings we have because of the pleasure of those who receive the benefits.

I said that, for Hume, natural and artificial virtues combine to produce actions and that sympathy plays a large role in this as well. I should add that the actions or utilities (or benefits) in question need not be ours. If the benefits that the actions produce are ours, we would approve the actions given our ability to sympathize in virtue of the fact that they are agreeable to us. If the benefits are not ours the principle of sympathy shines clearly as it enables us to receive by communication the sentiments of others as we approve of the actions that produce the benefits. The capacity to extend beyond or far out of ourselves and to feel the

pleasure or pain of others because of the agreeableness or disagreeable of the situation to them is explained by sympathy.[312]

4.3.3 Sympathy as an Appropriate Sentiment for the Problem of Secession

From the foregoing, it is clear that Hume consider sympathy as a natural sentiment in human relations. But how would Hume's principle of sympathy fare when applied to the problem of secession? Sympathy enables agents to commiserate, so to speak, with those in pain or those who suffer any form of hardship. We assume that it moves them to be supportive of those who suffer, especially when these conditions produce agreeable or disagreeable, pleasurable or painful feelings in them. The support they give in virtue of how they feel about the situation can be as extensive as possible; it can extend beyond just psychological support to physical and material support. In other words, sympathetic feelings enable individuals to be sensitive to the situations of others, namely, sympathy motivates people to support and promote the goals and life-plans of others.

Now, a theory of moral sentiments like that of Hume we suppose is able to accommodate the unendowed, less well-off or the unproductiveness of those from the North in virtue of the ability of agents to identify with or sympathize with others. The idea is that better-off members who happen to come from the South have the desire or sympathetic motivation to support the unproductiveness of those from the North under a theory of moral sentiments, even though they may not have the desire or rational motivation under a contractarian framework of reasons that is circumscribed by the maximization of expected utility. In virtue of the principle of

[312] Ibid, 3.3.1, pp.579.

resemblance and sympathy, those from the South come to actually experience passion and share in the affections of those from the North. They feel uneasy and pain—as those from the North do—because the unproductiveness and situation of the latter is harmful or disagreeable to them, i.e. those from the South. It is on the basis of these transferred affections that they are moved to support and cooperate with the North.

Given that sympathy is natural to all humans and given that we approve laws, social practices, and institutions because of their tendency to benefit the whole society, even where our interests are not at stake, the affections—pleasurable or painful feelings—we have for others are determined by the communication of the sentiments of these ones to us in virtue of agreeable or pleasurable and disagreeable or harmful situations. In these circumstances, supporting and cooperating with those from the North might decrease the utilities of those from the South. But since their moral compass and *weltanschauung* is neither grounded on the maximization of utilities nor on reasons, but on sympathy, they have no problem in accommodating and interacting with the North, even if that means a decrease in their utilities.

Hume's principle of sympathy thus aligns with Braybrooke's idea of a theory of moral sentiments and seems to support the view that in the foundations and applications of ethics, receptive sentiments can do the work that a contractarian framework of reasons, namely, Gauthier's theory of rational morality cannot do. If this is right, then Gauthier's view that Hume is at bottom a contractarian seems both misleading and mistaken. Hume's moral and political inquiries may lean themselves

toward scattered contractarian interpretations, but they are grounded on affectivities in their most fundamental aspects.

In expanding the horizon of the morality of sentiments, Braybrooke even supposed that it is possible to have a contract-based morality, one that spins on the wheel of affectivity. He calls it "a needs-based social contract", namely, "a social contract founded simply on upholding to begin with a principle of precedence for matters of need over matters of preference only." Although the contract of needs prioritizes needs over preferences it could still like Mb(CM)A be based on a deduction project, i.e. deduction of morality from self-interest, "but in every case it will be a deduction project governed by an unrefined conception of self-interest: one's property; one's liberties, opportunities, income, and power; one's needs."[313] Furthermore, the principles that might reconcile interests so conceived under a needs-based social contract "will be equally unrefined: uphold everybody's property rights; given equal liberties and fair equality of opportunity, distribute money income by the Difference Principle; heed anybody's preferences only insofar as it is consistent with meeting everybody's needs."[314]

The needs-based social contract infuses those results of self-interest with moral concern. But in this infusion, self-interest is "transcended or at least enlarged…. in which agents, in reckoning their self-interest, put a high value on having mutually sincere friendships."[315] An enlarged self-interest is an unrefined self-interest or a self-interest that is steered away from reason. The advantage of such enlarged, unrefined self-interest, according to Braybrooke, is that it

[313] Braybrooke, "Social Contract Theory's Fanciest Flight," *Ethics*, pp.762.
[314] Ibid, pp.762, 763.
[315] Ibid, pp.762-764.

incorporates sentiments or builds "on what is at least a close approach to moral concern." This is because "in accepting the terms of such a contract as they apply to other agents, every agent is not only accepting that the others' needs must be so far consulted if there is to be a contract. Every agent insofar as he or she is interested in an all-inclusive contract "is also in effect accepting what mutual concern for needs requires morally: that they take precedence over every agent's mere preferences, including her own."[316]

The moral gospel here from Braybrooke trumpets loud and clear. It is unmistakable in what it takes to be the appropriate ground for moral concern. In fashioning out an appropriate, just, and fair social contract *the needs* and not *the preferences* of agents is what matters or morally relevant. We proceed by consulting the needs of everyone. We consult the preferences of agents and factor them into the contract only if those preferences are consistent with everyone's needs, otherwise we label them antithetical to moral concern and moral progress, and throw them away through the window and into the 'bottomless pit.' Braybrooke's needs-based social contract thus builds on Hume's theory of moral sentiments insofar as it requires contractors to be sympathetic to the needs and not the preferences of other contractors.

Let me conclude by examining one issue about Hume's theory of moral sentiments, Mb(CM)A, and the problem of secession. Since Gauthier rejects Hume's views of moral sympathy, which he (Hume) founds on feelings that afford a direct identification with others, he cannot appropriate its elements or the totality

[316] Ibid, pp. 764.

of Hume's theory to defend issues of secession and moral progress.[317] In arguing for the Archimedean chooser, recall that Gauthier distances Mb(CM)A from the ideal sympathizer that is embodied in Hume's moral theory. In rejecting the ideal sympathizer as the appropriate embodiment of moral choice, Gauthier says, "we express our distance from Hume" and in distancing ourselves from Hume, "we agree with Kant that morality makes demands on us that are and must be quite independent of any fellow-feelings we may have."[318] Kant's moral framework and not Hume, and the former's "insistence that morality binds independently of the nature and content of our affections or feelings, Gauthier argues, is at least partially captured in the insistence that morality be based on the assumption of mutual unconcern."[319]

The Archimedean chooser is the liberal individual and the liberal individual is the ideal rational actor who possesses the affective capacity for (rational) morality and not the capacity for an affective morality. The former, identified with rational morality is consistent with the contractarian morality of reason, while the latter, identified with the ideal sympathizer is consistent with Hume's morality of sentiments. Now, since a rational morality rejects receptive sentiments it means it is unable to appropriate the resources of such morality in the test of application. Can

---

[317] Annette Baier pokes fun at Gauthier for once defending Hume as an ally in contractarian thought and appropriating his theory to his brand of contractarianism, then tuning around to label him an enemy of rational morality and contractarianism. She writes, "But [Hume's] theory relies on human 'affections', where Gauthier's project is to free us from bondage to them. Where once Gauthier had tried to appropriate Hume's theory to his own brand of contractarianism, now he cites Hume as mainly an opponent, and not an ally. Gauthier adopts Hume's instrumental and strategic conception of rationality, and like him outlines a series of progressively moralizing 'agreements', but joins Kant…in repudiating reliance even on fairly dependable human affections as any part of the basis for the morality that is to give us a free society," p. 315; Annette Baier's "Pilgrim's Progress," in *Canadian Journal of Philosophy*, vol. 18, 1988.
[318] MbA, p.238.
[319] Ibid, p.103.

we find within Mb(CM)A or a rational contractarian morality, for that mater, the resources to dissolve the problem of secession? Put in a different way, is it possible for a moral contractarian theory of reasons and refined self-interest, i.e. a theory that speaks to our considered preferences rather than our needs to solve the problem of secession? The answer to this question is provided in the following chapter, which I now turn to.

Chapter Five

Practical Reasons, Mb(DV)A, and the Problem of Secession

Introduction

One of the claims that was argued for in section 3.2.2 (Constrained Maximization and the Problem of Rational Compliance) is the idea that accepting the rationality of CM means a rejection of TRC's account of rationality. TRC treats practical reasons and rationality for that matter as strictly instrumental to the satisfaction of coherent considered preferences about outcomes, where the satisfaction of such preferences is explained in terms of the maximization of expected utility. We maximize expected utility, according to TRC, when we choose directly utility-maximizing actions. However, when we base our strategy on directly maximizing actions in strategic contexts we end up with less than optimal outcomes or fewer utilities. The implication of this is that TRC undermines its own foundation (the maximization of EU) which it seeks to promote.

Fundamentally, the CM account revises TRC's notion of practical rationality along the direction of dispositions. It identifies rationality with utility-maximization at the level of dispositions to choose. On this revised account of practical rationality (i.e. MRTC), individual preference is still basic to rationality, which is itself identified with utility-maximization. The modification is essential to Gauthier's overall strategy in Mb(CM)A. It enables him to advance the argument that it is rational for us to accept constraints on our utility-maximizing behavior in strategic contexts.

Mb(CM)A does not require an agent to choose directly utility-maximizing actions, but it requires that an agent choose dispositions, where those dispositions favor cooperation and allow that agent to maximize utility in some situation, given the strategies of those that he or she interacts with. An agent chooses and acts upon a disposition if and only if that agent expects to do better holding such a disposition than any alternative disposition. For example, if you and I have agreed to babysit for each other, even though by defecting, i.e. refusing to babysit for you I gain some additional utilities, I am *better off* performing my part of the bargain once you have performed yours or if I expect that you will perform yours. The reasoning seems intuitively appealing and it is this. You and I agreed to babysit for each other because both of us expect to do better forming the disposition to cooperate. If we did not expect the disposition to maximize benefits individually for us we would not have chosen it. Having formed the disposition it will be rational and individually beneficial for us to act on the disposition at the appropriate occasion.

By requiring us to choose directly utility-maximizing actions, TRC considers the probabilities of outcomes independent of actions. In contrast, by requiring us to choose dispositions, Mb(CM)A considers the probabilities of outcomes independent of actions but dependent upon dispositions. Mb(CM)A's identification of rationality with utility-maximization at the level of dispositions to choose is an improvement on TRC's conception of practical rationality. The advantage of this improvement is obvious. An agent that identifies with dispositions does better than one who identifies with directly utility-maximizing actions. This is because identification with dispositions provides an agent more

opportunities for engaging in cooperative activities, where those activities maximize that agent's utility profile. Yet, on both accounts of rationality the outcome (or EU) of an act (for TRC) or disposition (for Mb(CM)A) remains the product of the utility of its possible outcomes multiplied by the (subjective) probabilities of those outcomes. Both are EU-focused accounts, i.e. accounts that identify rationality with the maximization of expected utility.

As I indicated in chapter four, Mb(CM)A's view of practical reasons informs its solution to the problem of secession—the problem of what ought to be done with previously productive members of society who for some reason are now unable to contribute to the cooperative surplus. Given that a scheme of cooperation consisting of better-off or productive members and less well-off or unproductive members does not maximize the expected utility of better-off members, it is rational, according to Mb(CM)A, for better-off agents not to interact with or support less well-off agents. For Mb(CM)A *simpliciter*, when expected utility is stacked too high against cooperation, it is not rational to cooperate. I argued in that chapter that Mb(CM)A's solution to the problem of secession is a single-tracked silver bullet solution.

Mb(CM)A is a single-tracked silver bullet solution to the problem because it tracks only EU-reasons. By tracking only EU-reasons Mb(CM)A excludes from the horizon of rationality non-EU moral reasons. To track only EU-reasons is to offer a narrow and misleading characterization of practical rationality. My argument that Mb(CM)A fails in the test of application because it tracks only EU-reasons constitute in part an argument for the first part of Braybrooke's negative

thesis: the claim that Mb(CM)A cannot dissolve the problem of secession. Although I agree with the first part of Braybrooke's negative thesis, I reject the second part of the negative thesis, which is that any contract theory of reasons cannot solve the problem of secession. I am going to argue in this chapter that moral contractarianism is able to dissolve the problem of secession. The argument I shall be making accepts the general approach of Mb(CM)A, but modifies its conception of rationality along a decision-value (DV) direction.

Because the revision proposes a reading of Mb(CM)A that shows how situations of secession can be rational, it is essentially a replacement of Mb(CM)A by Mb(DV)A. I shall be arguing that unlike Mb(CM)A, which offers a single-tracked silver bullet solution to the problem of secession, Mb(DV)A offers a multi-tracked framework for solutions to the problem. Mb(DV)A identities rationality with the maximization of decision-value and tracks both EU and non-EU reasons (symbolic expressiveness and utilities), namely, the considered preference or aversion of agents for the acts that are available, in addition to the possible consequences of those acts.

In revising Mb(CM)A's conception of rationality, I shall follow Nozick's characterization of practical rationality in *The Nature of Rationality*. Nozick proposes a revision of decision theory along a DV direction. DV's modification is threefold. First, DV modifies decision theory by parsing EU into two parts—evidential expected utility and causal expected utility. Evidential expected utility (EEU) is the outcome probability given the act and causal expected utility (CEU), which is the outcome probability limited to what the agent can bring about in

choice situations. Second, EU is weighted not by the simple probabilities of the outcomes but by the conditional probabilities of the outcomes given the actions or a causal-probabilistic relation that indicates direct influence given also the actions. Third, DV recognizes the meaning or expressiveness acts have for EU-rational agents. This is the symbolic utility component of DV.

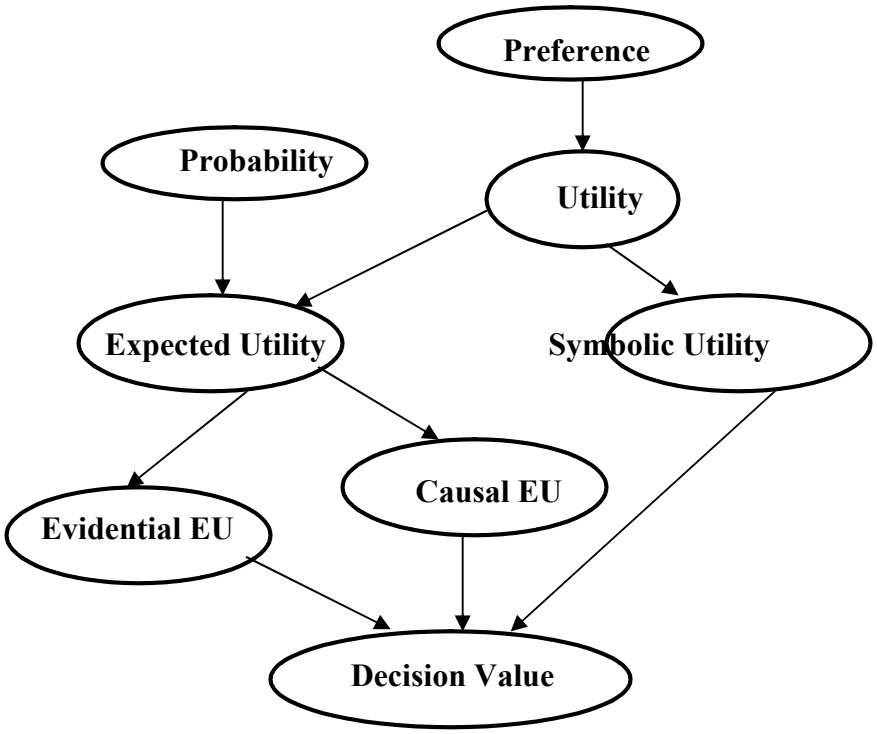
DV is act-sensitive as well as outcome-sensitive and it implies the rationality of taking into account utilities that attach to the acts other than the EU of the acts in an agent's payoff matrix. DV maximizes the weighted sum of two kinds of EU (EEU and CEU) and SU. The general picture of DV that emerges if we consider its three components, according to Wesley Cooper, is that "preference explains utility, and utility together with probability explains expected utility. Utility explains symbolic utility as a special case. Expected utility explains evidential expected utility and causal expected utility as special cases and these together with symbolic utility explains decision value."[320] This can be diagrammed as follows (see Figure 5.0).

There is at least one theoretical motivation for revising the theory of rational choice in a DV direction. The motivation for this revision is theoretically driven, namely: by the need to meet what one might call the idea of a general view of an acceptable and plausible account of rationality. This view claims that what determines the plausibility of an account of practical reasons or a description of rationality is the extent it takes into account all relevant factors and elements that play a significant role in an agent's choices. Because DV maximizes the weighted

[320] Wesley Cooper, "Nozick, Ramsey, and Symbolic Utility," *Utilitias*, Vol. 20, no. 3, September 2008, pp.304.

sum of EEU, CEU and SU, its description of rationality is superior to that of Mb(CM)A, which maximizes only EU. The latter is an EU-focused account of rationality because it tracks only EU reasons, but the former is a value-focused account of rationality because it tracks both EU and non-EU reasons. And because Mb(DV)A takes into account all relevant moral reasons and values (EU and non-EU reasons) that affect an agent's choices, it satisfies the idea of a general view of an acceptable and plausible account of rationality.

Figure 5.0: **The Three Components of a DV View** [321]



An application of the two kinds of EU would be helpful, particularly if we are to place in perspective the way Mb(DV)A provides a framework for solutions

[321] Ibid, p.305.

to the problem of secession. In the application of EEU and CEU, the focus is on how both components of EU factor into DV's resolution of long-standing paradoxes like Newcomb's Problem and the Prisoner Dilemma (PD). In contrast to Mb(CM)A, DV does not consider the probabilities of outcomes independent of actions. DV takes EU as weighted not by the simple probabilities of the outcomes but by the conditional probabilities of the outcomes given the actions (i.e. EEU) or a causal-probabilistic relation that indicates direct influence given also the actions (i.e. CEU). This is what I shall be demonstrating in the next section with the examples of Newcomb's Problem and the PD.

Following my examination of Newcomb's Problem and the PD, I shall examine in section 5.2 the difference between desire-based and value-based accounts of reasons for acting. I will argue that Mb(DV)A is a desire-and value-dependent account of practical reasons. The value-dependent nature of Mb(DV)A effectively distinguishes it from the theory of rational choice in general, which is a strictly desire-based account of practical reasons. In section 5.3, I put Mb(DV)A to the test of application, i.e. I examine how it provides a framework for solutions to the problem of secession. I conclude the section by examining the sense in which affective morality and Mb(CM)A are silver-bullet accounts.

5.1 Application of DV to Newcomb's Problem and the PD

Newcomb's Problem is well known and has been widely discussed after Nozick first presented and discussed it in "Newcomb's Problem and Two Principles."[322]

[322] Nozick's "Newcomb's Problem and Two Principles," in *Essays in Honor of C. G. Hempel*, N. Rescher et al (eds.), Dordrecht: Reidle, 1969, pp.114-146.

The problem is described as follows. There is a being who can predict your choices correctly. You have great confidence in the being's predictions. The being will either put $M or nothing in an opaque box (B2), depending on whether he predicts you are going to take both B2 and B1 (transparent box). If the predictor predicts you will take only B2 he puts $M in it. B1 has a significantly lower amount, say it ranges between a Loonie to $M minus a Loonie. You know about this, and the being knows you know that if he predicts prior to your choice that you will take only B2, he puts $M in it, otherwise he will put nothing in it. Furthermore, you are able to see a thousand dollars in B1. First, the predictor makes his prediction about your choice; then according to his prediction, he puts the $M in the opaque box or not; then you make your choice.

What should one do in this situation? The causal theory (i.e. CEU framework) recommends taking both boxes, which is the rational choice that reflects the 'dominant' choice because it ensures that one does best notwithstanding the being's action. The reasoning is that since the being has put $M in B2 or he hasn't, the only causal variable at play in this situation is one's choice. So one might as well take both the opaque and transparent boxes since one gets the extra amount that is in B1. In contrast, the evidential theory (i.e. EEU) recommends that one takes B2. Taking the opaque box is the rational choice that reflects conditional probabilities. It is a choice that is based on the evidence of the being's impressive prediction record. The idea is that since the being would certainly have predicted that one would take both boxes he would not put anything

in B2 and so one only gets the small amount in B1. Therefore, based on the being's impressive record, taking B2 would certainly guarantee one has the $M.

What makes this dilemma interesting is that its two horns consist of two arguments or solutions that pull in opposing directions. One solution recommends taking B2, while the other suggests taking both boxes. The causal theorist, grounding her reasoning on individual choice as the only casual variable at play, chooses the 'dominant' action and takes both boxes no matter the content of B1. Conversely, the evidential theorist, basing her reasoning on the impressive prediction record of the predictor, chooses the 'dominated' action and takes B2, regardless of how much or how little is in B1.

By parsing EU into EEU and CEU, the decision-maker is presented the choice of switching between purely probabilistic (taking only the opaque box) and causal/probabilistic reasoning (taking both boxes) depending on whether there is much to gain or lose by reasoning in terms of the latter or former. Parsing EU into EEU and CEU helps the decision-maker to avoid the problem of choosing between conditional probabilities and dominance. It allows the decision-maker to give more or less weight to either of the two choices, depending especially on the content of B1. If there is $M minus a Loonie in B1, since there is little to lose, *only a Loonie*, the decision-maker might as well give great weight to CEU. If there is only a Loonie in B1, one might assign great weight to EEU. There is little to gain by taking B1, *only a Loonie*.

DV's weighted sum of two kinds of EU does prove to be illuminating and perceptive for decision theory and helps to resolve Newcomb's Problem. Note that

DV's solution to Newcomb's Problem is not a single-tracked silver bullet solution; rather, it is a multi-tracked framework for solutions. DV does not instruct the decision-maker to take B2 or both boxes no matter the content or the value that is bestowed on the content of B1. On the contrary, it instructs the decision-maker to take B2 or both boxes depending on the content of B1 and the decision-maker's judgment of the role his or her choice or the predictor's prior record plays in the situation. A decision-maker switches to CEU depending on the content of B1 (a Loonie or $M minus a Loonie) and her estimation of the role of individual choice in the situation. She switches to EEU depending on the content of B1 (a Loonie or $M minus a Loonie) and her estimation of the impressive prediction record of the predictor.

The Prisoner Dilemma, which we encountered in previous chapters, is another long-standing problem that is congenial to an application of a DV account. I noted the following about the way the classic form of the PD is formulated: if one prisoner confesses and the other does not, the one that confesses walks away free (0 years in jail) and the second receives some years in jail (say 10 years of prison sentence). If both confess, each receives some years in jail (say 5 years). If both refuse to confess, each receives some jail time (say 2 years of prison sentence). The matrix below represents this information.

Figure 5.1a: **The Prisoner Dilemma with Matrix Showing Years**

	Prisoner 1	
	Don't confess	Confess
Don't confess (Cooperate)	2, 2	10, 0
Prisoner 2		
Confess (Rat)	0, 10	5, 5

In the above matrix the numbers are denoted in terms of prison years. I however, switch from this conventional numbering to utilities in figures 5.1b to 5.1f. In matrix 5.1a, the Prisoner Dilemma is reported in terms of literal prison years (0 for being cooperative, i.e. being free and 10 for being uncooperative, i.e. in the slammer for 10 years). But from matrices 5.1b to 5.1f I switched the convention. The matrices or numbers, i.e. 9, 2, 9.5, 4, 2.5, 7 etc do not literally stand for prison years but rather they stand for utilities, i.e. 9 or 9.5 are higher payoffs for the players for being cooperative (and which is equivalent to the 0 in matrix 5.1a, i.e. being free) and the smaller numbers, i.e. 2, 4 stand for lesser utilities or lower payoffs for the players for being uncooperative (and which is equivalent to the 10 in matrix 5.1a, i.e. being in the slammer for 10 years). So while 5.1a states the PD in literal prison years, matrices 5.1b to 5.1f express this in utilities. In 5.1a, 0 as per expected utility is better than being in the slammer for 10 years because 0 means being free; and in 5.1b to 5.1f, the bigger numbers as per expected utility is better than the smaller numbers because they stand for greater

utilities or higher payoffs. That is, being free has more instrumental value or is EU-superior to being in the slammer for 10 years. Matrices 5.1b to 5.1f express this instrumental value and valence in terms of bigger numbers/greater utilities or higher payoffs.

As with any Prisoner Dilemma (PD), the cooperative optimal outcome diverges from the equilibrium suboptimal outcome; in addition, the cooperative solution payoffs are higher than the ratting solution payoffs. As figure 5.1b illustrates, the cooperative solution payoffs are 7, 7 utilities while the ratting solution payoffs are 4, 4. Since each decision-maker aims to maximize her utilities one would expect that the rational choice is for either to cooperate; cooperation draws higher utilities (3 more) than confessing. However, as the dilemma unfolds, both decision-makers choose to rat rather than to cooperate.

Figure 5.1b: **The Prisoner Dilemma with Matrix Showing Utilities 1**

	Prisoner 1	
	Don't confess	Confess
Don't confess (Cooperate)	7, 7	2, 9
Prisoner 2		
Confess (Rat)	9, 2	4, 4

There are four possible payoffs in the matrix boxes for each decision-maker. The structure of the dilemma, in its classic form, is that of cooperation (not

confessing) and non-cooperation (ratting or confessing), and the 'dominant' choice or action is for each person to rat no matter what the other person does. As we saw with Mb(CM)A's solution to the PD, CM allows decision-makers to avoid the equilibrium suboptimal outcome (end up with 4 utilities). We are disposed to CM and we sufficiently know that others are disposed to CM, and the reason we cooperate with others is that we expect to do better by disposing ourselves to keep our commitment, and the disposition provides us greater utilities. However, as we saw in chapter three, there are reasons to be skeptical of the fruitfulness of dispositions not just as a solution to the PD but in Gauthier's theory in general.

The PD parallels Newcomb's Problem in at least one relevant sense—both involve two solutions that pull in opposing directions. The solutions to Newcomb's Problem are either (a) take B2 or (b) both boxes. The solutions to the PD are either (a) rat or (b) cooperate. The argument to rat is based upon the dominance principle congenial to CEU, while the argument to cooperate is based upon the optimal principle congenial to EEU. The CEU solution to the PD recommends performing the dominant action by ratting when one thinks that causal probabilities represent causal influence or if one sufficiently ascribes pessimistic probabilities to the other's chance of cooperating. The EEU solution to the PD recommends performing the cooperative action when one believes that the other person is relevantly similar to one. EEU favors cooperation based on one's expectation that the other person will do as one does, even though one's action does not causally affect what the other does. The conditional probabilities here do not represent any casual influence. The probabilities of outcomes are conditional exclusively upon

what the decision-maker can make happen in the choice situation. Even though the decision-maker's choice is 'mirrored' by the choice of the other decision-maker in some relevant sense, it does not mean that the former's choice causes the latter to choose in specific ways.

A different way of characterizing the sense in which the weight that decision-makers give to each of the particular principles of CEU and EEU encourages a shift in choice or strategy in the PD is to pay more attention to the numerical utilities entries in the matrix. As we saw in figures 5.1a and 5.1b, confessing is the dominant choice and it also appears to be the 'rational choice.' We might suppose that this is precisely because there exist significant differences among the payoffs. But once we close or narrow the gap among the payoffs, what we take to be the rational choice appears to change. What will happen if the matrix entries are changed as illustrated in figures 5.1c and 5.1d?

Figure 5.1c: **The Prisoner Dilemma with Matrix Showing Utilities 2**

		Prisoner 1
	Don't confess	Confess
Don't confess (Cooperate)	**8.5, 8.5**	2, 9
Prisoner 2		
Confess (Rat)	9, 2	**2.5, 2.5**

Figure 5.1d: **The Prisoner Dilemma with Matrix Showing Utilities 3**

	Prisoner 1	
	Don't confess	Confess
Don't confess (Cooperate)	**7.5, 7.7**	4, 9.5
Prisoner 2		
Confess (Rat)	9.5, 4	**7, 7**

In figure 5.1c cooperation appears to be the rational choice, and in figure 5.1d ratting seems to be the rational choice. The reason for the difference in the rational choice in both figures is due to the difference in the numerical entries (payoffs). In general, when the cooperative (EEU) solution payoffs are significantly higher than the ratting (dominance or CEU) solution payoffs, as in figure 5.1c, cooperation seems rational and the dominance argument has little force. On the contrary, when the cooperative solution payoffs are only slightly higher or better than the ratting solution payoffs, as in figure 5.1d, ratting appears rational and the dominance argument has much more force.

So far, I have limited my analysis of the PD to cases where acts are taken to be merely instrumental to outcomes that are desired by a decision-maker. I now want to examine cases where the PD includes a decision-maker's considered preference or aversion for the acts that are available, that is, the symbolic expressiveness and utilities of the acts themselves (SU) or the intrinsic valence of

the acts independent of their instrumental value. To do this would require that we do not take the utilities in the payoff matrix to be strictly expected utility.

Recall the example of Jonas and the wallet in chapter four. If Jonas returns the wallet to lost and found, he does so not simply because of expected utility calculations, but because of his considered preference for the act of retuning it or his considered aversion for the act of keeping it. If he keeps the wallet, he does so not merely because of the possible outcomes of the acts, but because of his considered preference for the act of keeping it or his considered aversion for the act of returning it. The same is true here in the Prisoner Dilemma What this means is that in addition to the instrumental value of the acts, we need to consider as well the act's meaning, or symbolization, or expressiveness. Specifically, we consider the considered preference or aversion of the decision-makers to (a) the act of ratting and (b) the act of cooperating, in addition, of course, to the possible outcomes of the acts.

If we suppose that the act of ratting expresses something of value for a decision-maker, then this expressiveness will factor into her decision whether or not to perform the act, taking into account as well the expected utility of the act. However, if we suppose that a decision-maker is averse to the act of ratting because of what the act means to her, then she would not perform the act. Alternatively, if we suppose that a decision-maker is averse to the act of cooperation because of what the act means to her, then she would not perform the act. However, if we suppose that cooperation expresses something of value for a decision-maker then this expressiveness will factor into her decision whether or not to perform the act,

taking into account as well the expected utility of the act. Another way of stating this is to say that if, for a decision-maker, the act of ratting means betrayal or the act of keeping quiet expresses the value of cooperation, then we may suppose that in choosing the 'optimal' action of doing what is best for both collectively, instead of the 'dominant' action of doing what is best for one person, no matter what the other decides, the decision-maker employs the action to represent her or his value as a cooperative person.

I have indicated that SU is the third element in Nozick's DV account of practical rationality and one of his additions to decision theory. DV, as we have seen, factors the utility profile of the agent into two distinct elements: EU, (which is outcome-sensitive) and SU (which is act-sensitive). Typically, SU differs from EEU and from CEU because Nozick takes it as an additional subjective element and broadly uses it to express some belief or value we hold as rational agents, or to say something positive about what kind of individuals we are, or want to be. SU assigns a kind of psychological utility to acts that, by their nature from the agent's point of view, symbolize a whole class of like actions, or have an expressive quality, or otherwise mean something, quite apart from the tendency of those acts to cause certain outcomes.[323]

Certainly, this characterization of SU and the way it is related to actions in virtue of expressiveness is insightful and fruitful in shedding light on the nature of rationality. By factoring SU into the matrix, we take into cognizance a decision-makers' preference or aversion for the acts, such that that decision-maker may choose to cooperate when cooperation draws slight losses or punishment (say, a

[323] Nozick, *The Nature of Rationality*, pp. 27,43.

few more weeks in jail if the other decision-maker chooses to rat) and the SU of cooperation is sufficiently high to outweigh the losses. But a decision-maker may choose not to cooperate when cooperation draws significant losses or punishment (say, 5 more years in jail if the other decision-maker chooses to rat) and the SU of cooperation is sufficiently low to outweigh the losses.

Suppose a decision-maker has preference for the act of cooperating. Suppose we can represent this in terms of utilities, say, **5** (i.e. **SU** of being cooperative). Although noncooperation is still the 'dominant choice' that yields the best payoff whatever the other player does, adding **5** to the payoff for cooperation makes not ratting the rational choice. I represent this in figure 5.1e, a version of figure 5.1b, but this time with an SU component. However, I should point out that in a real payoff matrix with numbers, SU does not appear. This is because SU is not an EU or a function of an action's outcomes. It is represented in the payoff boxes as a way of illustrating its place and role in rational choice.

Figure 5.1e: **The Prisoner Dilemma with a Symbolic Utility Component 1**

	Prisoner 1	
	Don't confess	Confess
Don't confess (Cooperate)	7 (+ **5 SU**), 7 (+ **5 SU**)	2, 9
Prisoner 2		
Confess (Rat)	9, 2	4, 4

If we take out the **5 SU** from the payoffs boxes, which we should, but factored into the utility profile of the decision-makers, we have the following as is illustrated in figure 5.1f.

Figure 5.1f: **The Prisoner Dilemma with a Symbolic Utility Component 2**

		Prisoner 1
	Don't confess	Confess
Don't confess (Cooperate)	**12, 12**	2, 9
Prisoner 2		
Confess (Rat)	9, 2	4, 4

Taking into account utilities other than the EU in the payoff matrix in the PD represents a decision-maker's choice to embrace the cooperative solution rather than the dominance solution, and hence the shift from figure 5.1b to 5.1e or figure 5.1f. In figure 5.1b, the decision-maker chooses the dominance solution (which offers her **4** utilities) in the absence of SU, even though that meant a loss of **3** utilities. However, this changes significantly in figure 5.1f as soon as SU is factored in. In figure 5.1f, the decision-maker chooses the optimal or cooperative solution, which has a higher utility of **12**.

Like we saw in the case of DV's solution to Newcomb's Problem, DV's solution to PD is not a single-tracked silver bullet solution but a multi-tracked framework for solutions. DV stipulates that the decision-maker picks a solution

(cooperating or ratting) based on the relevant weight that decision-maker has bestowed upon each solution. The recommendation (á la CEU) is that we perform the dominant action by ratting when we believe that causal probabilities represent causal influence or if we sufficiently ascribe pessimistic probabilities to the other's chance of cooperating. But, if we think that the other person is relevantly similar to us, the recommendation (á la EEU) is that we perform the cooperative action. DV provides the same framework for solutions to the PD if we approach it from a DV/SU perspective. We are told to cooperate just in case cooperation draws slight losses or punishment and the SU of cooperation is sufficiently high to outweigh the losses; otherwise, we are told to choose the dominant action and not cooperate. In other words, DV/SU recommends we perform the cooperative act if that act symbolizes for us some value and it recommends that we rat just in case the cooperative act does not symbolize for us any such value.

5.2 Practical Reasons Explained by a Desire-and Value-dependent Mb(DV)A Account

In this section, I shall be examining desire-based accounts and value-based accounts of reasons for acting. At the end of the section, I hope to have argued the following: (1) Mb(DV)A is a desire-and value-dependent account of practical reasons and (2) Mb(DV)A explains practical reasons better than the theory of rational choice in general and Mb(CM)A) in particular. I begin the first half of the section (5.2.1) with a discussion of the difference between desire-based and value-based reasons for acting. In the latter half of the section (5.2.2), I examine the sense

in which we might think of Mb(DV)A as a desire-and value-dependent account of practical reasons.

5.2.1 Desire-based Accounts and Value-based Accounts of Reasons for Acting

To begin let me summarize the definition of utility that we first encountered in chapter three. Utility is the measure of our relative satisfaction from consuming a variety of baskets of goods. Utility captures individual preferences in the sense that it is a measure of our coherent considered preferences about outcomes, and as such, it remains subjective. Utility is relative to our desires and beliefs, and thereby is sensitive to context. We may specify utility along two dimensions or connections. Utility may be specified along symbolic connections or along causal or probabilistic connections. When utility is specified along symbolic connections the utility of a symbolized situation is imputed back to the action. And utility is specified along causal connections when such utility is not imputed back to the action. Utility, in the second sense, together with probability, explains expected utility (see figure 5.0).

In the theory of rational choice, utility is specified in the second sense, and when this is the case desires or preferences are taken to provide reasons for acting. This is a desire-based or preference-based account of reasons for acting, according to which the reasons for acting are provided by certain facts about how we could fulfill or achieve our present desires, preferences or aims.[324] Some of these desires

[324] Henceforth, when I use 'desire' it should be taken to mean not just what we want, or want to achieve, i.e. our aim but also preference. Desire-based (or aim-based or preference-based) accounts are rooted in rational choice theory's overarching view of rationality. As was clearly evident in chapter three, rational choice theory is an EU-focused account of practical reasons. It is the view

may be what we actually *want* to achieve or they may be what we would now have if we were rational, namely, thinking clearly, aware of the relevant facts, and had gone through some rigorous process of rational and informed deliberation.

Suppose I am now hungry and have the desire to eat sushi, then on a desire-based account, I ought to choose those acts that would enable me fulfill this desire. Given my desire for sushi, I ought to go to a Japanese restaurant rather than a Mexican restaurant or to a wine tasting event, and I ought to ask the waitress to serve me sushi rather than Japanese noodles or miso soup. My reasons for acting are provided by the facts that these acts would enable me fulfill my desire for sushi. The reason for my choosing these acts is explained by my desire for sushi and my reason for acting on the desire is provided by these acts, which fulfill the desire. I would not go to a Japanese restaurant if not for my desire for sushi and I would not ask to be served sushi if not for my desire for sushi.

This way of putting it is a bit misleading for it suggests that the reason for acting, for all desire-based accounts is provided by the relevant desire or those acts that would enable an agent to fulfill a particular desire. Some desire-based accounts, according to Derek Parfit, are a little more sophisticated in the way they specify desire-based or desire-dependent reasons.[325] A 'naïve desire-based account' claims that the reasons for acting on a desire are provided by the relevant desire or the act or acts that would enable us to fulfill the particular desire. But a 'sophisticated desire-based account' claims that the reasons for acting on a desire are provide by facts outside the desire.

---

that as rational agents, we seek to maximize EU, and we choose the best action or strategy according to our stable preferences and the constraints we face.

[325] See Derek Parfit, *Climbing the Mountain* (new book manuscript), p.28.

A sophisticated desire-based account argues that although the reasons for acting *depend* on the relevant desire, they are not *provided* by it or by the fact that having a particular desire would give us some reason for trying to fulfill that desire. Rather, the reasons for acting on a desire are provided by certain other facts, most of which causally depend on our having the desire. If your desire for sushi gives you happiness or some gustative pleasure, and assuming that the non-satisfaction of this desire would leave you distressed, then, on a sophisticated desire-based account, what gives you reason for trying to fulfill the desire is the gustative pleasure you derive from consuming sushi.

Note that, on a sophisticated desire-based account, we locate the reasons for acting on the desire for sushi outside the desire itself or those acts that enable us to fulfill that desire. We locate the reasons for acting on the desire in the pleasure we would get from fulfilling the desire. The reason for my going to a Japanese restaurant and not a Mexican restaurant, and the reason for my asking the waitress to serve me sushi rather than Japanese noodles is explained by my desire for sushi, and my desire for sushi is explained by the pleasure I derive from fulfilling the desire.

The problem with these accounts (naïve and sophisticated) is that the reasons for acting are still desired-based reasons in the sense that practical reasons are mainly reasons for acting. Practical reasons mainly are provided by facts about what would fulfill or achieve our present desires. But this is completely misleading and as Parfit opines, quite rightly in my view, practical reasons are not merely or mainly reasons for acting. We also have reasons to choose those acts that fulfill our

desires or "to *have* the desires or aims that our acts are intended to fulfill or achieve."[326] Whereas Parfit wants to extend these reason-giving facts to include the rationality of doing one's duty, or doing what one morally ought to do, even when one does not want to do that, I wish to limit the reason-giving facts to the rationality of choosing acts that one believes are valuable, or that symbolize something about one, in addition to the possible outcomes of those acts.

We might want to reject desired-based accounts that take practical reasons to be merely reasons for acting on the ground that they present a misleading picture about the reasons for acting on desires. My desire for sushi is given by the facts that also give me reasons for acting and I would have these reasons even if I did not have the desire for sushi. That I know that certain acts would fulfill my desire for sushi or that I am aware that I will get certain pleasure from fulfilling the desire for sushi does not tell the complete story about my reason for acting on the desire for sushi. Moreover, to claim that the reasons for acting are provided by certain facts, most of which causally depend on our having certain desires, suggests that we ought to act on a particular desire insofar as we are aware of the fact that it would enable us to fulfill that desire. For example, if I have the desire to count all the blades of grass in the lawns at the University of Alberta and if I know the facts that realize this desire, say, the fulfillment of this desire provides me some gustatory pleasure, and since I want this pleasure, I should go and start counting the blades of grass.

The speciousness of naïve and sophisticated accounts can clearly be illustrated with the following example. Consider someone—let us call her

[326] Ibid, p.28.

Tammy—who has the desire for 'no-talent-development.' This desire we may say is essentially a preference to be a 'couch potato.'[327] Tammy is aware of the facts that would enable her to fulfill this desire—she knows that certain various leisure activities provides her pleasure—so she performs these acts. It could be said that for every person that has this desire there would be a number of acts that would give him or her pleasure, i.e. fulfill the desire. In the case of Tammy, if we suppose that counting some or all the blades of grass in her city and watching most of the reality and sports programs on TV are what provide her pleasure, then, according to naïve and sophisticated accounts of reasons for acting, she ought to engage in these activities.

One might say that there is a sense in which Tammy's desire violates Kant's categorical imperative. As a couch potato, she operates on a maxim that conflicts with rational willing, namely, an imperfect duty requiring one to develop one's talents. In Kant's moral ontology rational beings have an imperfect duty to develop their talents or powers in ways that would make them rational and meaningful members of the kingdom of ends or, shall we say, in ways that would make them productive and useful rational beings.[328] For Kant therefore, Tammy

[327] I am equating the desire for no-talent-development with the preference to be a couch potato. We might think of them as identical in the sense that both involve doing nothing or, rather, both involve what Kant seems to refer to as 'a life of indulgence', i.e. a life of engagement in leisure activities that provide one with pleasure. See Kant's *Groundwork of the Metaphysics of Morals*, 4:423. Even though there may be some sense in which the desire for no-talent-development and the preference to be a couch potato are relevantly different, I will be using them interchangeably in my discussion.

[328] It does seem deceptive to say that Kant characterizes the lack of talent development in terms of usefulness or productivity. Rather than usefulness, he characterizes it in terms of one's duty and the conflict of rational willing. He argues that developing one's talents is an imperfect duty. Therefore, not developing one's talent violates one's imperfect duty. It violates one's duty because the maxim that arises from the action results in a conflict of the will or rational willing. See Kant's *Groundwork of the Metaphysics of Morals*, 4:423. Nonetheless, there is a sense in which we may interpret the view that an agent would not desire to live in a world of no-talent-development as a

compromises her rational nature when she acts on the desire for 'no-talent-development.'

Because Kant's moral ontology is framed in absolutist terms he can certainly make this evaluation. However, naïve and sophisticated accounts of reasons for acting cannot. This is because the desire that Tammy acts on and her reason for acting seem consistent with what these accounts take to be practical reasons. If what provides an individual reason for acting on a desire is the desire itself or the pleasure that individual derives from fulfilling that desire, then insofar as Tammy derives pleasure fulfilling the desire for 'no-talent-development,' she acts invariantly with what naïve and sophisticated accounts take to be practical reasons.

It might be objected that Tammy's desire to be a couch potato is not 'rational' because it thwarts her other desires, namely, desires that she presently has or those she hypothetically would have in the future. If it is the case that the desire to be a couch potato does not promote her future wellbeing, then it does seem irrational for Tammy or anyone to form such a desire. This is a valid objection that can be raised by a desire-based theorist who accepts some form of desire-based account of wellbeing.

On a desire-based account of wellbeing what counts as reason or reasons for acting is our rationally chosen end, i.e. our future wellbeing. We limit the range of opportunities or activities available to us by forming and acting on the desire to

---

general view of usefulness. 'It is my duty to develop my talents as a lawmaking member of the kingdom of ends' becomes 'it is my duty to be a useful member of the community of legislators.'

be a couch potato.[329] Since the range of activities available to us is essential to achieving our well-informed ends, our desires can be said to rational only if they promote or contribute to these ends.

A different way of putting it is to say that our wellbeing is the external standpoint that provides us the basis for having a particular desire or for trying to fulfill a particular desire. And what makes a particular desire rational is that the desire does not (1) thwart the desire for wellbeing or our ends and (2) conflict with other desires that lead to the desire for wellbeing or our ends. Let us call this view the 'connection thesis' because it aims to connect desires with wellbeing or our ends. Specifically, the 'connection thesis' takes a particular desire to be dependent on the desire for wellbeing, that is my desire for sushi is dependent on my desire for wellbeing.

One might reject the 'connection thesis' on the ground that not all of our desires connect to other desires or the desire for wellbeing. What makes this objection plausible is that wellbeing is neither a thing nor a quality of a thing we possess whose quantity can be measured. Because wellbeing is not measureable, it is difficult to say precisely how or when a particular desire thwarts it. In principle, some desires may connect to an individual's life going well overall. An example might be the desire to eat often. We need to eat regularly to be able to carry out our

[329] This is a self-interested or largely instrumentally based argument, and it differs from the duty-based argument that Kant advances, which is neither grounded on benefits nor *necessarily* related to the interest of the agent. On some instrumentally based view, the reason why we develop our talent is the benefits we get from doing so. But on a duty-based account, we are required to develop our talent because it is our duty to do so, notwithstanding that when we develop our talent we produce benefits which we share in. If it is the case that 'my duty to develop my talents as a lawmaking member of the kingdom of ends' becomes 'it is my duty to be a useful member of the community of legislators', then usefulness might be desired for its own sake and not for the benefits I get.

normal human functions; hence, we might say that the desire to eat contributes to our life going well overall or our desire for wellbeing.

However, to say that fulfilling the desire to eat often contributes to our wellbeing or to say that the desire to eat regularly connects to the desire for our life going well overall is not to say that all of our desires connect to our desire for wellbeing. It is difficult to see how, for example, a person's desire to play a game of chess or solitaire, or to count the blades of grass in the lawn, or to color a picture in a coloring book, in other words, to engage in some 'innocent delights' to use Locke's catchword connect to that person's life going well. Even when we suppose that the desire is about something as important as the desire to eat, it is difficult to see how every choice of food is connected to a person's life going well. When you reach for the piece of candy in front of you this one time or other times, for example, it is preposterous to claim that the reason you reached for the candy is that you believe it would make your life go well. It is possible that you reached for the piece of candy because you think it contributes to your happiness, but it is conceivable also that you took the candy for any number of reasons that are unconnected with your belief or desire about your life going well.

Beyond this, desires cannot be self-supporting. Suppose that if after informed rational deliberation, Tammy wants future wellbeing or happiness as an end. This desire might provide her instrumental reasons to have some other desires. That is, the desire for future wellbeing might presumably give her reasons to want whatever would contribute to her wellbeing. But the fact that she has the desire for future wellbeing can hardly be defended as what provides her a reason to have the

desire. Her desire for wellbeing as an end cannot truly give her a reason to want wellbeing as an end.[330]

In any case, the claim that the desire for future wellbeing gives an individual reasons to want whatever would contribute to that individual's wellbeing seems to support the argument that, on desire-based accounts of wellbeing, we ought to try to fulfill a particular desire insofar we know the fact that would enable us to fulfill that desire. In Tammy's case, she clearly knows the fact that would enable her to fulfill the desire for 'no-talent-development'—activities that provide her with pleasure. If she is happy acting on the desire, then there is a sense in which we might suppose that her wellbeing is promoted when she watches TV and counts blades of grass.

If Tammy is happy fulfilling the desire, then it is not quite clear why a desire-based account of wellbeing would consider her desire to be a couch potato irrational. Tammy's desire may *be* irrational. Indeed, her acting on or trying to fulfill the desire may be irrational, but it seems to me that this is not an evaluation that a desire-based theorist who accepts some form of desire-based account of wellbeing can make.

The last point can be made more forcefully by exploring the connection between Tammy's desire and beliefs. In particular, if we assume that her desire to be a couch potato is explained by the belief she holds regarding the sort of natural abilities she possesses one might be able to say that a desire-based account of wellbeing is committed to the view that her desire is rational. Consider the following as Tammy's beliefs for having the desire to be a couch potato.

[330] See Parfit, *Climbing the Mountain*, p.38.

(a) The belief that she lacks productive natural abilities.

(b) The belief that she would be happy if she idles away her time.

Without going into issues of coherency of beliefs or issues as to whether both of Tammy's beliefs satisfy the norms of rational belief, we can understand the sense in which the beliefs explain her desire not to develop her talents. If she believes (a) that developing her talents is a waste of her time given the sorts of abilities she has and (b) that she would be happier idling her time away, then it will seem rational, on a desire-based account about wellbeing, for her to act on the desire not to develop her talents. Although this proves that particular desires are linked to particular beliefs it is important to note that it does not necessarily follow that it is the beliefs that provide reasons for acting. One might accept the view that desires are explained by beliefs without been committed to the view that beliefs provide reasons for acting. It is possible that beliefs explain desires and ends, but it does not follow that they provide reasons for acting.

When one claims that beliefs do not provide reasons for acting, even though they might explain desires, what one is claiming is that it is not sufficient to point to facts about the beliefs as the reasons for acting on the beliefs or desires. If I decide to run away or climb a tree when confronted on a lonely path by a grizzly bear, my reason for running away or climbing a tree, one might say, is not provided by the facts about my belief about grizzly bears, namely, that grizzly bears are dangerous or that the bear will attack me if I do not run away or climb a tree. It

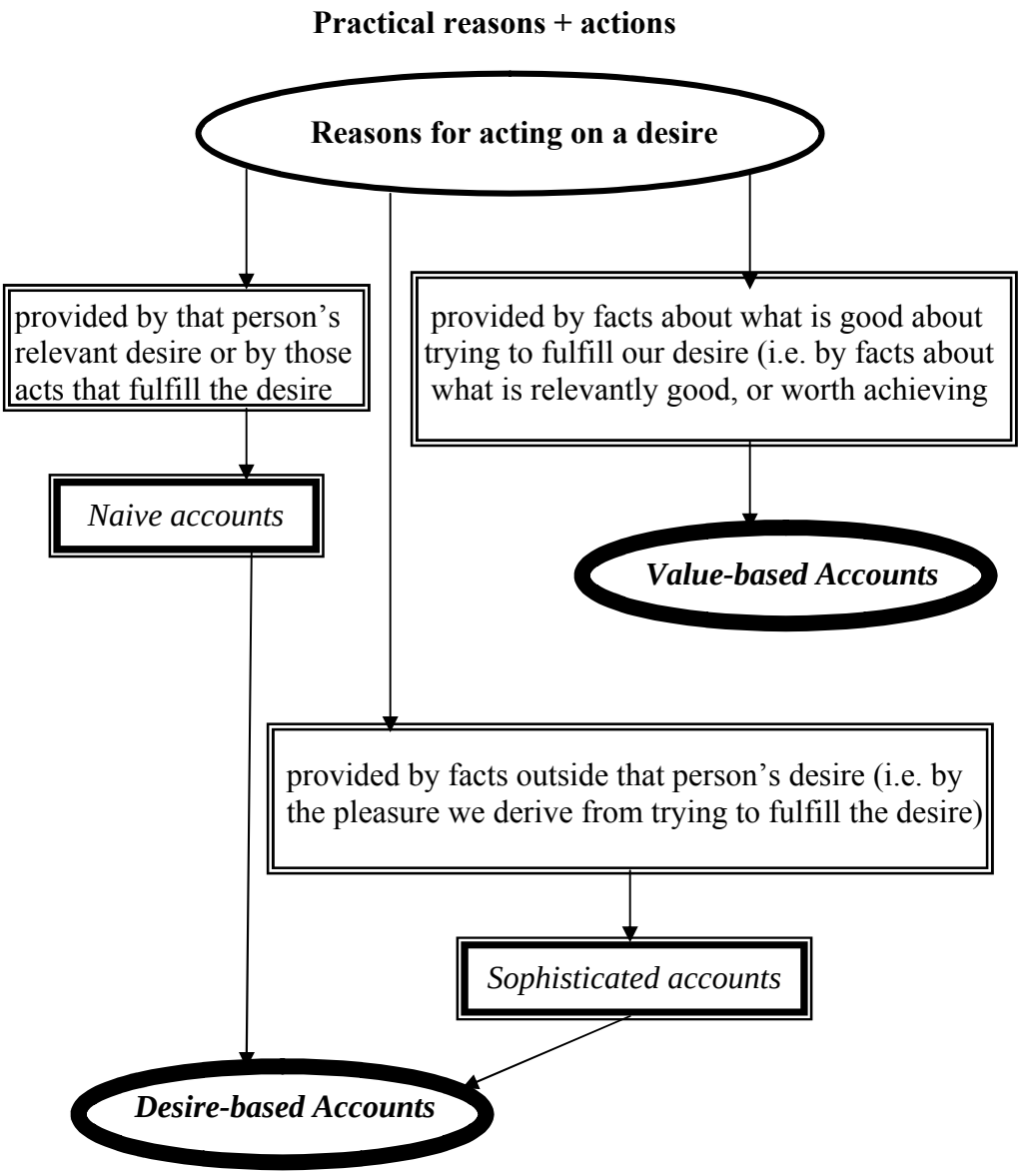
seems rather, that my reason for running away or climbing a tree is provided by some value, shall we say, the value of what life is to me or means to me.

Rather than evaluate practical reasons from the prism of desire-based accounts, we can evaluate reasons for acting from the standpoint of value-based accounts. In general, value-based accounts claim that practical reasons are neither merely reasons for acting nor are they provided by facts about what would fulfill or achieve our present desires; rather, they are provided by facts about what is good about trying to fulfill our desires. The reasons for acting on a desire are 'provided by the facts that make certain possible outcomes, actions worth producing or preventing, or make certain things worth doing for their own sake.'[331] If I want future wellbeing as an end or if I desire certain acts, my desire for future wellbeing or those acts cannot be defended as what provide me reasons for having the desire or trying to fulfill it. My desire for future wellbeing as an end or those acts is provided by something else, by facts that make having future wellbeing as an end worth having.

Naïve and sophisticated desire-based accounts argue that practical reasons are merely reasons for acting, but as we have seen, there are problems with taking practical reasons as simply reasons for acting. In contrast, value-based accounts argue that practical reasons are not merely reasons for acting, but are provided by facts that make our desires, actions or outcomes worth achieving. The representation in figure 5.2.1a highlights the primary area of difference between reasons for acting on value-based accounts and reasons for acting on naïve and sophisticated desire-based accounts.

[331] Ibid, p.28.

Figure 5.2.1a: **Reasons for Acting on Desire-based and Value-Based Accounts**



As the above illustration demonstrates, the reason-giving facts for fulfilling a desire, on naïve and sophisticated desire-based accounts, are not provided by the goodness or badness of the desire. The reasons for acting are also not provided by the value of the acts associated with the desire or the outcomes that fulfilling the desire bring about. Rather, they are provided by the desire itself or by those facts

that fulfill the desire. By contrast, the reason-giving facts for fulfilling a desire, on value-based accounts of practical reasons, are provided by facts about what is good about trying to fulfill the desire.

As the name suggests, value-based accounts are organized around the idea of values. That is to say, for any value-based account, values are decisive as reasons for acting. There are descriptive as well as normative dimensions to any value-based account. The descriptive dimension concerns how we, in fact, *do* act. For any given person, i.e. *p*, then values (i.e. facts that make us want to fulfill a desire or achieve an aim) are decisive for *p*, in the sense that they provide reasons for acting, for *p*. The normative dimension concerns how we *ought* to act. For any given person, i.e. *p*, then values (i.e. facts that make us want to fulfill a desire or achieve an aim) ought to be decisive for *p*, in the sense that they *must* provide reasons for acting, for *p*.

I should hasten to add that there is one further element in the normative aspect of value-based accounts. Call this the 'meta-normative' aspect of value-based reasons. A value-based account is meta-normative when it evaluates a normative claim or counter normative claims about facts that ought to make us want to fulfill a desire or achieve an aim. A normative claim of reasons for acting on a desire advances facts that ought to make us want to fulfill that desire. A meta-normative evaluation does not make a claim *per se* but evaluates normative claims or counter claims about facts that ought to make us want to fulfill a desire or achieve an aim. It evaluates a normative claim for acting on a desire when it evaluates facts—put forward by a value-based account—that ought to make us

want to fulfill that desire, and it evaluates counter normative claims for acting on a desire when it evaluates rival claims—put forward by value-based accounts—about facts that ought to make us want to fulfill that desire.

What kind of facts—which ought to make us want to fulfill a desire—will a value-based account for the reason for acting put forward for, say, the desire for 'no-talent-development'? If one accepts the claim that value-based accounts provide some deep insights into the reasons for acting, insights that naïve and sophisticated desire-based accounts as well as desire-based accounts of wellbeing cannot provide, then the facts that ought to make one want to fulfill the desire for 'no-talent-development' can neither be located in the acts that fulfill the desire nor on the pleasure one derives from acting on the desire, but on what is relevantly good about the desire. Let us suppose that a value-based account takes the reason for acting on the desire for 'no-talent-development' to be 'provided by such fact as 'a simple lifestyle,' which may or may not be related to the simple lifestyle of the monk or nun, who practices religious asceticism. If a simple lifestyle is what is relevantly good about trying to fulfill the desire, the question that we should be asking is "why is a simple lifestyle good?"[332] I explore speculatively this question in the rest of the section.

[332] Parfit may object to this way of specifying value-based accounts, i.e. specifying, on a, value-based account, the reasons for acting on the desire for no-talent-development in terms of 'a simple lifestyle. His objection may come from the fact that he takes the rationality of doing one's duty as what is relevantly good about trying to produce certain outcomes or achieve certain aims. But since an individual, say, the ascetic (not the religious ascetic, i.e. monk or nun) who values a simple lifestyle may value it because he or she considers it relevantly good and because he or she considers it relevantly good, one might say that it satisfies Parfit's claim that what is relevantly good about trying to produce certain outcomes or achieve an aim should not diverge from what one morally ought to do. And since I limit the reason-giving facts to the rationality of choosing acts that one believes are valuable or that symbolize something about one, in addition to the possible outcomes of

The question what is good about a simple lifestyle invites if nothing else a meta-normative evaluation, an evaluation that takes into account both an intrinsic and a non-intrinsic justification for what is good about a simple lifestyle. A value-based account provides an intrinsic justification for the desire to be a couch potato if that account takes a simple lifestyle to be good for its own sake, and a value-based account provides a non-intrinsic justification for the desire for couch potato just in case that account appeals to something else to cash out what is good about a simple lifestyle. In either case, the meta-normative evaluation is abstract in the sense that what is being evaluated is the fact or facts about what is good about trying to fulfill the desire for 'no-talent-development.' If we assume that a value-based account provides an intrinsic justification, say, it takes a simple lifestyle as what is relevantly good about trying to fulfill the desire for 'no-talent-development,' and if it takes a simple life as good for its own sake, then what a meta-normative evaluation is primarily concerned with is whether a simple lifestyle is good for its own sake.

Even though the focus and interest of Karl Marx lies elsewhere, his view of the relationship between work and humanity is certainly one that can be exploited by a meta-normative evaluation of whether a simple lifestyle is good for its own sake. Work, according to Marx, defines our humanity.[333] If we suppose that this is right, then on a meta-normative evaluation, a simple lifestyle is neither good because of something else nor for its own sake on the ground that it destroys one's

---

those acts, achieving a simple lifestyle is relevantly good insofar it has value for one or symbolizes something about one.

[333] I understand work here narrowly, as a productive economic activity in which one exerts strength or faculties to achieve something.

humanity. The point being made is this: by failing to engage effectively and appropriately with the world through work, Tammy or anyone who refuses to develop her talent abases or destroys her humanity.

Alternatively, a meta-normative evaluation might consider whether a simple lifestyle is good for its own sake on the following grounds:

> **P1**: a world that has many more activities is better than a world that has fewer activities.
> **P2**: talents are necessary for a world of many more activities, both to create such as a world and to effectively appropriate it, i.e. make us of it.
> **Conclusion 1:** from P1 – P2, one that has no talents or that does not develop his or her talents cannot effectively appropriate a world of many activities
> **P3**: a life that engages in as many activities as possible is a worthwhile one.
> **P4**: One that lives a simple lifestyle does not engage in many activities.
> **Conclusion 2:** from P3 – P4, a simple lifestyle is not worth achieving since it is a life of fewer activities and insofar it is not worth achieving it is not good for its own sake.

If Tammy accepts the reasoning underlying the above arguments, then she accepts that it is irrational to form the desire for 'no-talent-development.' The logic of the arguments *qua* meta-normative evaluations suggests what sorts of desires Tammy ought to have and act on. It is rational to form a desire if that desire is relevantly good, where what is relevantly good may be specified instrumentally or intrinsically. Now, there are reasons why someone might reject the reasoning underlying the above arguments or some, if not all, of the premises of the arguments. This is not the place for me to consider them.

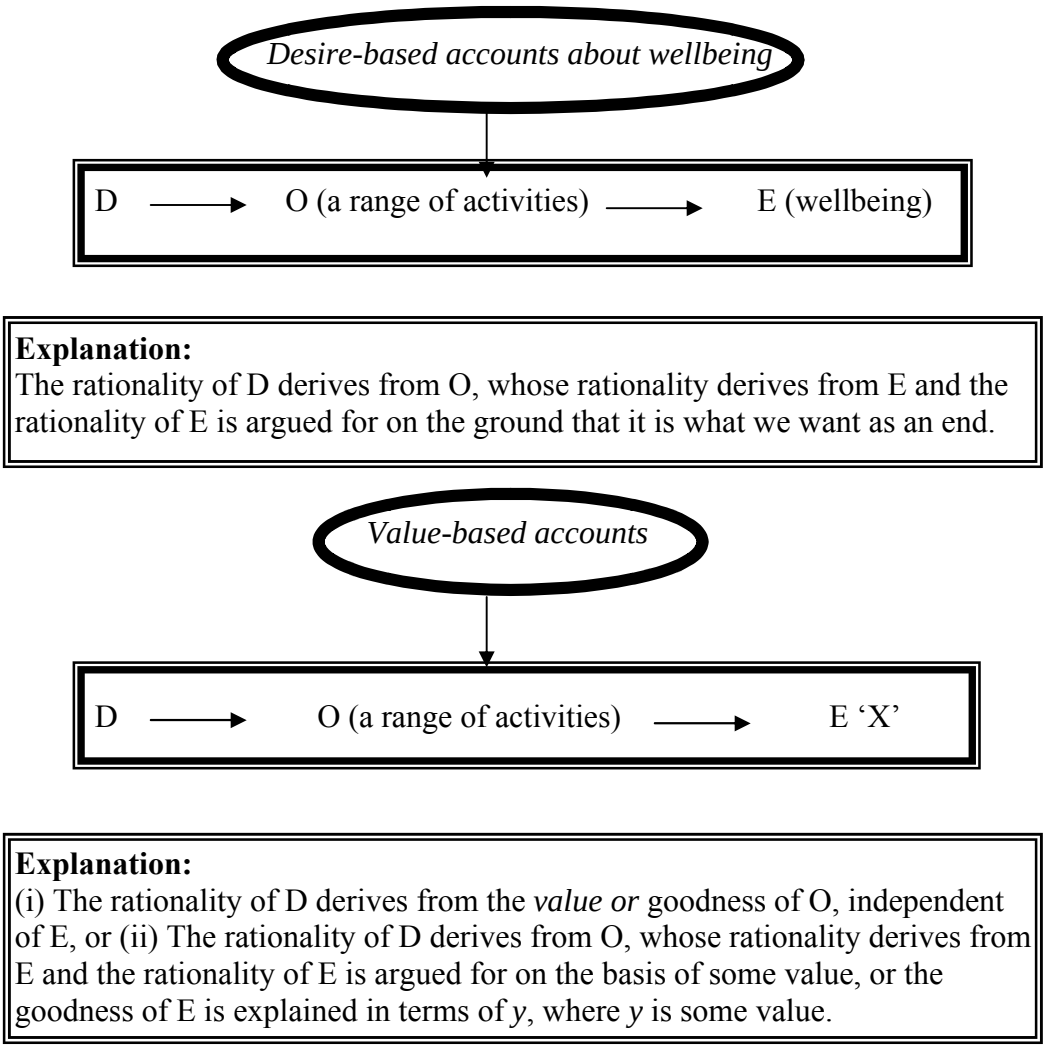
However, it is important to point out that a person who believes in the value of a simple lifestyle might accept the claim that talents are useful and necessary to create a world of many more activities, in which case that person might accept some of the premises of the arguments. Yet, that person might reject the conclusion that a simple lifestyle is not good for its own sake. Tammy might reject the conclusion that a simple lifestyle is not good for its own sake on the ground that a simple lifestyle is all she needs to get by or going in a world of many activities.[334] Tammy's rejection of the conclusion does not question the claim that the sorts of or range of activities a person engages in are for the most part related to that person's abilities. It also does not question the view that the more abilities a person has the more activities that person can engage in. Rather, it questions the claim that because a simple lifestyle is a life of fewer activities such a lifestyle has less or no value or is not worth achieving.

Let me emphasize that a value-based account about the goodness of desires, outcomes or ends, or about what is good about trying to fulfill a particular desire, or trying to achieve a certain aim or end, is as problematic as any desired-based account of wellbeing that seeks to examine the best perspective from which to judge the good life. What I have done thus far is to present an outline of how naïve and sophisticated desire-based accounts of reasons for acting *qua* desire-based accounts of wellbeing and a value-based account of practical reasons cash out the desire for 'no-talent-development.' It might be interesting, for theoretical purposes, to see how desire-based accounts of wellbeing and value-based accounts of the

[334] Note that the content of a simple lifestyle is not the same for Tammy and the ascetic, be it the non-religious ascetic or the religious ascetic, i.e. the monk and the nun.

reasons for acting cash out the corresponding desire, i.e. the desire for 'talent-development.'

Figure 5.2.1b: **Desire for Talent-development + Justifying Reasons**



Note that, as figure 5.2.1b illustrates, although both desire-based accounts of wellbeing and value-based accounts of the desire for 'talent-development' sketch a close connection between the desire and a world of a range of activities, they do not cash out the relationship the same way. If **D** is the desire for 'talent-development'; **O** is the outcome of the desire (a world of many more activities

made possible by developing one's talents or the state of affairs that holds in virtue of our acting on **D**) and; **E** is what **O** enables us achieve, say, an end-goal, then we get the following representation for the desire for 'talent-development' on both desire-based accounts of wellbeing and value-based accounts about reasons for acting.

I should point out that the application of the above representation is limited to certain types of desire-based accounts of wellbeing and value-based accounts of reasons for acting. Specifically, it is limited to desire-based accounts of wellbeing and value-based accounts of reasons for acting that cash out **O** in terms of a range of activities and those that respectively specify **E** as either wellbeing or "X" (i.e. unknown). Of course, on a complete representation, **O** is unknown on both accounts. On the above representation, **E** is not unknown, on desire-based accounts of wellbeing, even though on a complete representation **E** is unknown. **E** is unknown because it can be specified in any number of ways. Note also, that **E** for value-based accounts is unknown. The reason for this is that even when value-based accounts specify **O** in terms of a range of activities, **O** can be instrumental or intrinsic.

O is instrumental when its rationality derives from what it enables an individual to achieve, in this case **E**. **E** on this representation can be identified with the "good life" or specified in any number of other ways (see explanation (ii) of figure 5.2.1b). **O** is intrinsic if it is valuable for its sake, in which case its rationality does not derive from anything else, namely, there is no content for **E**. When there is no content for **E** **O**'s rationality can be derive either from **O** or from

something else. This something else could be anything from the meaning associated with engaging in a range of activities or some particular value (see explanation (i) of figure 5.2.1b).

To summarize, naïve and sophisticated desire-based accounts about reasons for acting take practical reasons to be merely reasons for acting. On the latter's account, the reasons for acting are provided by the relevant desires or those acts that fulfill the desires; e.g., the reason for your buying sushi is explained by your having the desire for sushi. On a sophisticated account, although the reasons for acting depend on the relevant desires, they are not provided by the desires; rather, they are provided by other facts; e.g., the reason for your buying sushi is because of the pleasure you derive from consuming sushi. In contrast to desire-based accounts, the reasons for acting according to value-based accounts of practical reasons are provided neither by the desires nor by the fact that fulfill them. The reason-giving facts are provided by facts about what is motivatingly good about trying to fulfill the desires; e.g., the reason for your buying sushi has to do with what is relevantly good about the desire for sushi, or what is relevantly good about the outcome from eating sushi.

5.2.2 DV/SU *qua* Mb(DV)A as a Desire-and Value-Dependent Account of Practical Reasons

In section 5.1, I discussed how a DV account of practical rationality offers a multi-tracked framework for solutions to Newcomb's Problem and the PD, two lingering paradoxes that have been extensively discussed in the literature. By parsing EU

into EEU and CEU, DV demonstrates that the weights assigned to the solutions in both problems significantly affect the solution that is selected. The weighted DV account (DV/EEU-CEU) is value-sensitive because the agent engages the various solutions of the problems by considering what is relevantly good about each solution, i.e. the value of a Loonie versus the value of $M; the estimation of the predictor's impressive record vis-à-vis that of the role of individual choice in the situation; the difference between spending a few more weeks in prison and five or more years in the slammer.

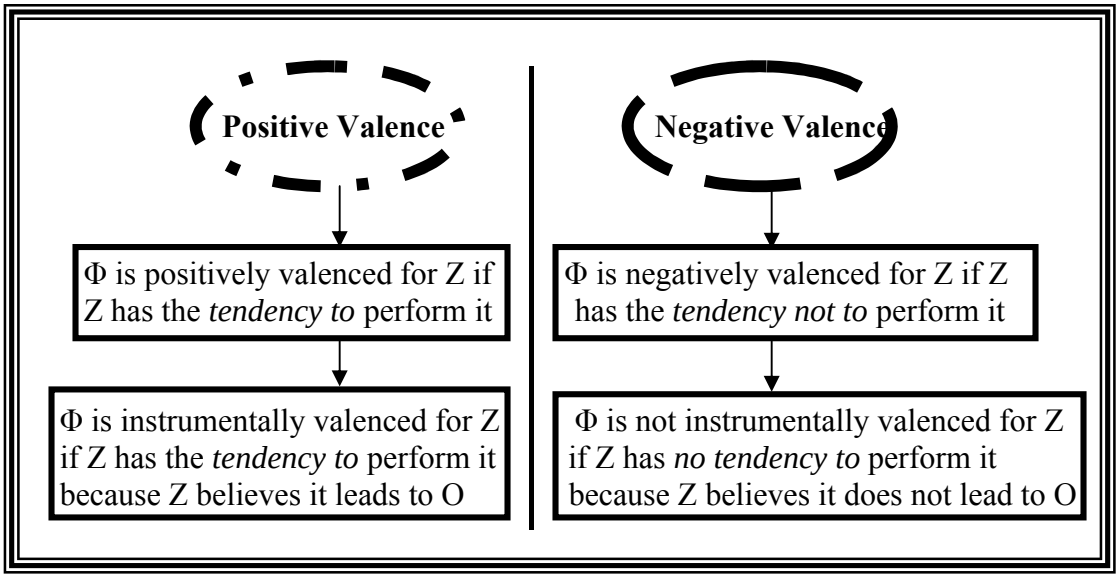
DV/EEU-CEU offers an alternative to the single-weight or single-track EU account of rational choice, of which Mb(CM)A is a species. However, DV/EEU-CEU is still EU-focused and consequence-dependent. The decision-maker in the PD who chooses cooperation when cooperation draws only slight losses does so because of the possible outcomes of the act—it is not rational ratting the other person when cooperation offers only a small gain, i.e. draws only a few more weeks in prison. By contrasts, SU, the third component of DV, is action-dependent and sensitive; it is about the meaning or expressiveness of acts, i.e. the utility that acts may have intrinsically for their own sake.[335] However, like DV/EEU-CEU, DV/SU is multi-track. Because DV/SU *qua* Mb(DV)A is action-sensitive, it, and not DV/EEU-CEU provides the multi-tracked framework for solutions to the problem of secession that I will be discussing in the next section. What I want to do

[335] We might represent SU as arising from Intrinsic Utility (IU), that is to say, as a species of the genus IU, where IU *qua* SU is different from EU. Whereas, EU is a function of the probability of an act leading to some outcome, IU is not. On this account, SU is not value-free since it refers to an agent's valenced for an act because of its perceived value (goodness or badness). The perceived value may have to do with the act's meaning, its representativeness, or its expressiveness.

in this section is to explain how DV/SU *qua* Mb(DV)A fits with the value-based accounts I have been discussing. Specifically, I want to show how it is a desire-and value-dependent account of practical reasons.

To explain the relationship between a decision-value account and value-based accounts of reasons for acting let me begin by elaborating on the features of actions and valence that we briefly encountered in section 5.1. Let Z stand for any given agent; O for any possible outcome for that agent; and Φ (Phi) for the act that produced the outcome. For Z, positive and negative valence can be represented as illustrated in figure 5.2.2c.

Figure 5.2.2c: **Positive and Negative Valence of Φ for an Agent (Z)**



From figure 5.2.2c, the following conclusion can be drawn. Φ is positively *intrinsically* valenced for Z just in case there is something motivationally good about it, i.e. Φ has positive intrinsic utility for Z. Φ has positive intrinsic utility for

Z if the valence of Φ is not dependent on Z's beliefs that it brings about O or additional goals that Z wants. To say that Φ has positive intrinsic utility for Z is to say that Z has the tendency to perform Φ or will perform Φ. Such positive intrinsic utility or valance for Φ is because of Φ's perceived value—it is valuable to act compassionately or honorably, to be there for a friend, to avoid lying, and so on.

DV/SU's multi-tracked account stands out in cases where it is clear that EU does not sufficiently explain why people perform certain acts. Consider this:

1. Person F decides to go to prison rather than give up an associate.
2. Person E prefers the more painful of two ordeals.

The single-tracked EU account of rational choice does not explain why F would choose to go to prison rather than give up a member of a group. Presumably, not going to prison maximizes F's EU, and going to prison does not maximize F's EU. Let us say that the $\mathbf{EU^0}$ (0 EU) and $\mathbf{EU^1}$ (1 EU) are respectively the possible outcomes from not going to prison and going to prison. Yet, F chooses to go to prison, even though the act has 0 EU. As well, the single-tracked EU account of rational choice does not tell us why E prefers the more painful of two ordeals. Supposedly, the less painful ordeal and not the more painful ordeal provides E greater expected utility. Yet, E prefers the more painful ordeal.

By contrasts, DV/SU explains why F might choose to go to prison rather than give up an associate; it also tells us why E might choose the more painful ordeal. DV/SU explains that going to prison rather than betraying an associate for F and preferring the more painful ordeal for E are respectively intrinsically valenced for F and E, and may be rational (though they may not be as well, depending on the

weights or values that F and E assign to the acts). Those who prefer not to betray an associate, or prefer a more painful of two ordeals, do not *necessarily* do so because of EU-benefits. Their decision may always be because this experience or ordeal, as Parfit rightly puts it, 'has some other feature, such as being deserved, or it presents the occasion to show how tough they are'[336], namely, it is valuable to act honorably or to be loyal. On the interpretation of a DV/SU account, expected utility does not completely explain the reasons why people would prefer to go to prison or the more painful of two ordeals.

It is important to emphasize that value, for DV/SU, is not objective but subjective. SU is subjective because it is determined by the agent's considered preferences rather than an objective ideal. An account of value is objective if it claims that values, although connected to our preferences and beliefs, lie outside of us and exist independently of our determination .[337] And it is subjective if it claims that values are not independent of our determination. To think of values as objective is to regard them as existing independently of our determination and as providing a standard to govern our desires. A DV/SU *qua* Mb(DV)A account of subjective value is compatible with the account of subjective value that Gauthier defends in Mb(CM)A.

If we say that Φ_1 stands for a Toronto Blue Jays' game (T), and Φ_2 stands for a Boston Red Sox's game (B), then instrumental and intrinsic valence as they relate to an individual's subjective determination can be illustrated below.

[336] Parfit, *Climbing the Mountain*, p.36.
[337] An example of a value objectivist is Plato. Plato regards beauty, justice, virtues, truth and good as objective that we may have more or less clear knowledge of.

Figure 5.2.2d: **Instrumental and Intrinsic Valence for T and B**

For T

1. Φ_1 is instrumentally valenced for you if you have a tendency to perform it because you believe it leads to O.
2. Φ_1 is not instrumentally valenced for you; its valence is extinguished.

Therefore, you do not perform Φ_1.

And

1. Φ_1 is positively *intrinsically* valenced for you just in case it has positive intrinsic utility for you.
2. Φ_1 does not depend on your belief that it leads to O or to further goals that you want.

Therefore, you perform Φ_1.

For B

3. Φ_2 is instrumentally valenced for you if you have a tendency to perform it because you believe it leads to O.
4. Φ_2 is not instrumentally valenced for you; its valence is extinguished.

Therefore, you do not perform Φ_2.

And

3. Φ_2 is positively *intrinsically* valenced for you just in case it has positive intrinsic utility for you.
4. Φ_2 does not depend on your belief that it leads to O or to further goals that you want.

Therefore, you perform Φ_2.

There is at least one motivation for rejecting an objective account of value. The claim that value exists independently of our determinations is a claim that would appear to threaten the thesis of individualism. We encountered this thesis in earlier chapters, mainly in connection with our discussion of the demand of mutual advantage. The thesis simply states that we may neither collapse a person's conception of the good with those of others nor compel anyone to accept the principles of social relationships. That is to say, the relevant factor in evaluating our willing compliance with moral and social demands is that we individually determine the moral principles that are operational in society. Because a subjective account of value claims that individuals *must determine* what counts as values and reasons for acting, it respects or meets the demand of the thesis of individualism.

Indeed, value-based reasons may originate outside of us, in the sense that they may connect to something external to us, say to our culture or to our social upbringing, but we draw on them and they enter into our rational deliberation when the actions we choose are determined by reasons we take to be valuable. Of course, it is specious to claim that the values we associate with all come from us individually. We belong to societies and by nature, we are social, namely, born into customs and traditions of our own particular society. Much of what is significant about us *qua* human beings is the consequence of our upbringing and social context. As Nozick rightly puts it, we live in a cultural world, a world that is richly imbued with values and symbolic meanings and we draw on this when we act.[338]

However, although we draw on value-based reasons or symbolic meanings when we act, we have to determine them for the reasons to count as subjective. We

[338] Nozick, *The Nature of Rationality,* p.32.

determine value-based reasons when we appropriate them as our own, and we appropriate them when we evaluate them as valuable from our standpoint. We agree with the view, endorsed by Parfit and Kant in some form, that acting presupposes a prior conception of moral principles. We cannot be motivated to act unless we recognize some prior obligations or values. If one is to be moved to act by a sense of justice or duty, one must antecedently believe some action to be just or one's duty. However, the antecedent duty one acts upon cannot be grounded in some external ideal independent of our subjective determination.

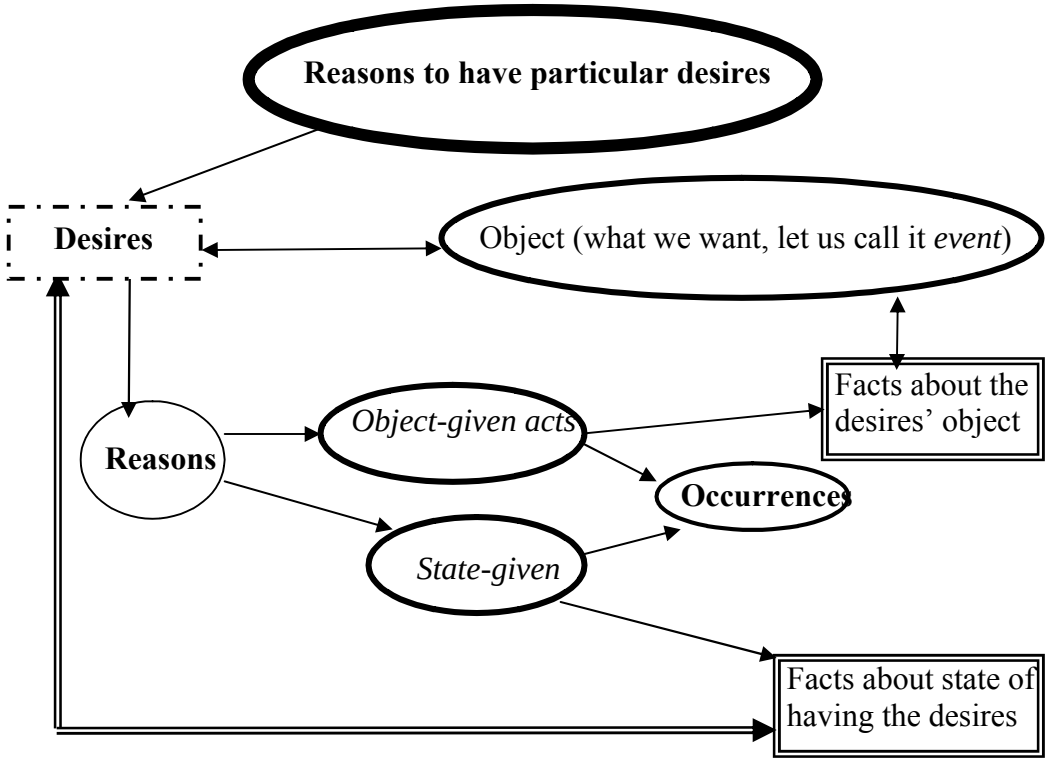
I now want to explicate how DV/SU connects to the value-based accounts that Parfit offers. In particular, I want to show how DV/SU is a desire-and value-dependent account of practical reasons, i.e. a variant of the desire-based view with value coloration or a variant of desire-based account that is sensitive to value or with an emphasis on value. In what follows, I adopt the sketch that Parfit provides[339] regarding the ways in which value-based accounts claim we can have reasons to have particular desires. All desires, according to Parfit, are intentional. By this, he means that they have objects, which are what we want. If you like, we can call these objects events, in the wide sense that they cover states of affairs or outcomes, and acts. Parfit parses the reasons for desires into two, as I illustrate in figure 5.2.2d.

Of our reasons to have some desire, some are *object-given*, while others are *state-given.* Reasons are state-given when they are provided by certain facts about our state of having the desire and they are object-given when they are provided by

[339] Parfit, *Climbing the Mountain*, pp. 31, 32.

certain facts about the desire's objects or events.[340] According to Parfit, these object-given and state-given reasons can take three forms.

Figure 5.2.2e: **The Relationship of Object-given and State-given Reasons to Desires**

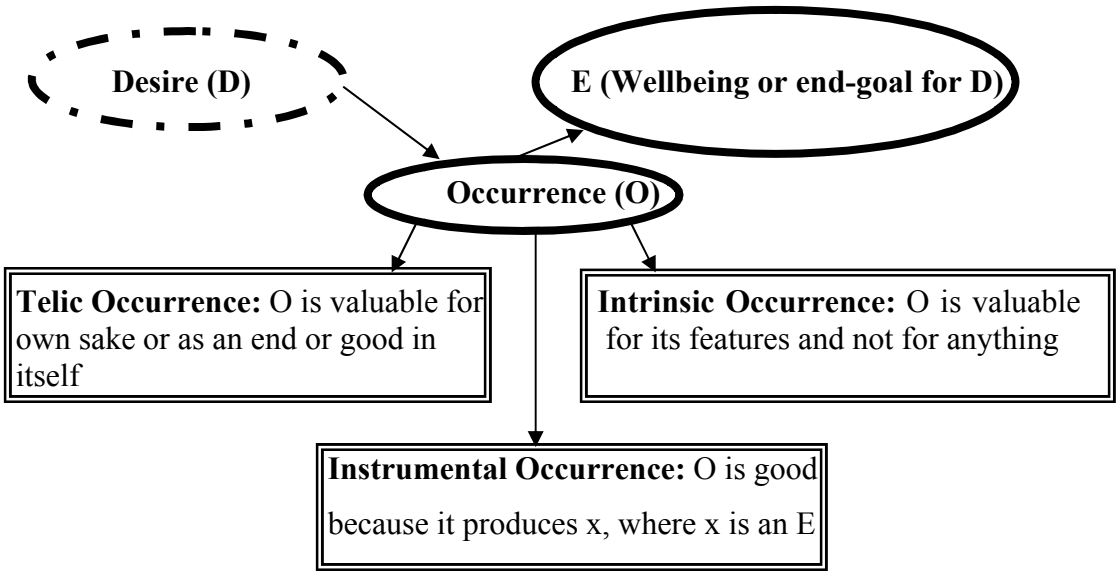


If we call the object-given and state-given reasons 'occurrences,' then the table of 'reasons for' *event* would be as represented in figure 5.2.2e. Occurrences can be *telic*, when they are provided by facts that make the *event* good as an end, or worth achieving for its own sake. Occurrences can be *intrinsic*, when they are provided by the *event's* intrinsic properties or features. And occurrences can be *instrumental*, when the reason for our wanting some *event* is because the *event* will help produce, or be a means of achieving, some good end.[341]

[340] Ibid, p.31.
[341] Ibid, pp.31, 32.

Figure 5.2.2f: **Reasons and Occurrences on Value-based Accounts**

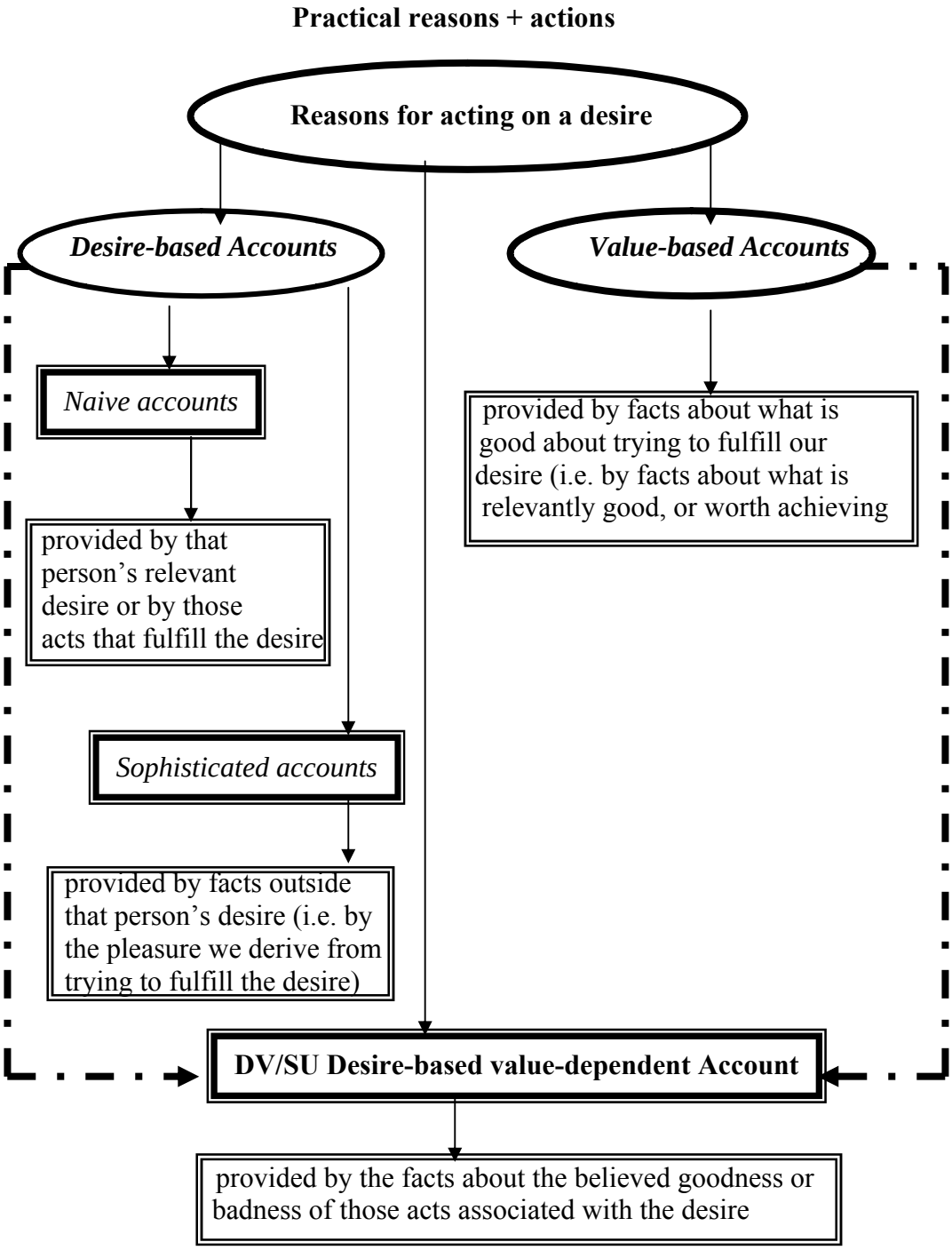
Table of reasons for any *Occurrence* for Value-based accounts



DV/SU is about the meaning or expressiveness of acts; an act's positive intrinsic valence, not for its instrumental value, but for its intrinsic features or value, and to this extent it shares much in common with telic and intrinsic occurrences. SU is utility that an act may have intrinsically for its own sake, or the utility that attaches to acts, or our belief about those acts independent of our belief about the outcomes of those acts or further goals that we may desire. Telic occurrence claims that some of our reasons are provided by facts that make what we want good as an end or worth achieving for its own sake. Intrinsic occurrence claims that the intrinsic features of what we want provide some of our reasons for acting. DV/SU claims that some of our reasons are provided by the believed goodness or badness of those acts or by the meaning or expressiveness of those

acts. On both telic and intrinsic occurrences as well as on DV/SU, the goodness of an event is independent of its instrumental value.

Figure 5.2.2g: **DV/SU and Desire-based/Value-based Accounts**



In what sense is DV/SU, then, either a variant of the desire-based view with value coloration or a variant of desire-based account with an emphasis on value? From our discussion of reasons for acting, it is clear that what provides reasons for acting, on both desire-based accounts and value-based accounts, are certain facts, i.e. certain facts about how we could fulfill, or achieve, or ought to make us want to achieve our desires or aims. What sets them apart is primarily what these facts are. figure 5.2.2f illustrates these differences.

A naïve desire-based account says these facts are the relevant desires or those acts that fulfill the desires. To the question, "why should I fulfill my desire for sushi?" a naïve desire-based account provides this answer: because I have the desire for sushi. So my desire for sushi is what gives me reason for trying to fulfill the desire for sushi. Or my desire for sushi is what gives me reasons for choosing those acts that fulfill the desire for sushi. A sophisticated desire-based account associates these facts with what we get fulfilling the desire, e.g., pleasure. To the question, "why should I fulfill the desire for sushi?" a sophisticated desire-based account responds this way: because of what I get from consuming sushi (pleasure, happiness, etc). So the expectation of happiness or the pleasure I will derive from eating sushi is what gives me reason for trying to fulfill the desire for sushi or for choosing those acts that would fulfill the desire I have for sushi.

But Parfit rightly asks, why should we want the desire or pleasure in the first place? This question is a question about what reason or reasons we have for wanting these things (desires, or the actions that fulfill the desires, or pleasures, or outcomes). For him, these facts or the reasons for wanting these facts are the

rationality of doing one's duty, or doing what one morally ought to do, even when one does not want to do that. To the question, "why should I fulfill the desire for sushi?" Parfit's value-based account responds in this manner: because of what is relevantly good about the desire for sushi or pleasure, or about tying to fulfill the desire, or wanting what the desire gives us (e.g. pleasure). So what is relevantly good about my desire for sushi or what is relevantly good about the pleasure that I get from eating sushi is what gives me reason for trying to fulfill the desire for sushi. Or, the rationality of doing what is morally required or good is what gives me reasons for choosing those acts that would fulfill the desire for sushi.

Since DV/SU ties rationality to maximization of utility—utility understood in DV terms—it is committed by the definition of utility to a desire-based view—a desire-based view that is value-sensitive. To the question, "why should I fulfill the desire for sushi?" a DV/SU desire-and value-dependent account responds as follows: because the desire or the acts associated with the desire are motivationally good. So, what makes me want to or try to fulfill the desire for sushi is the believed goodness or badness of those acts associated with the desire. Or, what gives me reasons for choosing those acts that would fulfill the desire for sushi is that they are positively intrinsically valenced for me.

On this account, DV/SU is a variant of the desire-based view for the reason that it is committed by the definition of utility to a desire-based view. But it has the elements of the value-based view in the sense that the value of those acts is what gives an agent reason to choose them. An agent chooses the acts that fulfill a desire—or an agent wants to fulfill a desire—not solely because of the outcome of

those acts or the outcomes associated with fulfilling the desire (e.g. pleasure, EU) but because the acts are motivationally good, i.e. the acts have certain value for the agent or they symbolize something for him or her. While a DV/SU desire-and value-dependent account takes the value of those acts, explained in terms of the maximization of DV (SU and EU) to provide reasons for acting, Parfit's value-based account takes the rationality of doing one's duty, or doing what one morally ought to do as providing reasons for acting.

There is something eerily Suitsian and Kojevian about telic and intrinsic occurrences. The view that the goodness of an event or what we want is independent of its instrumental value is similar to the non-instrumental 'for its own sake,' or 'for the sake of nothing' standpoint of Bernard Suits' Grasshopper utopic vision of "games played for their own sake."[342] According to the Grasshopper's vision, the only prime activity left after the elimination of all instrumental activities—eliminated due to the existence of plenitude and abundance—is playing games. The games in the Grasshopper's vision have no instrumental value; they are played not for the outcome they bring about, but for their own sake.

The non-instrumental 'for its own sake,' or 'for the sake of nothing' outlook of telic and intrinsic occurrences also remind us of Alexandre Kojeve's view of the future of history. The future of history is something that interests Kojeve,[343] who, in interpreting Hegel's view of history, suggests that since humans are constituted by their labor, they will die and become dandies at the end of

[342] See Bernard Herbert Suits's *The Grasshopper: Games, Life and Utopia*, Toronto, Broadview Press, 2005. Original edition published in 1978.
[343] See pp.160, 388, note on Alexandre Kojeve's *Introduction to the Reading of Hegel*, Allan Bloom (ed.) and trans. by J. H. Nichols, Jr., New York, Basic Books, 1969.

history. As dandies, there will be nothing left for them to do, no inventions or paradigms to create, no discoveries, revolutions, and science to excite them. Post-historical humans will be dandies and all their productive labors would resemble Grasshopper's game-like activities, where such activities are instrumentally empty and accomplish nothing.

As a variant of the desire-based view that is value-sensitive, DV/SU recognizes the fact that an act may be intrinsically positively valenced on the basis of its intrinsic features, even though its instrumental value is believed nil. To the extent that DV/SU recognizes the utility that attaches to acts, or the belief about the acts, independent of the acts' further outcomes, it expands on the instrumental account of rationality that Gauthier presents in Mb(CM)A. However, in so doing, it replaces Mb(CM)A conception of practical rationality, which is EU-specific with a broader and more robust conception of practical rationality that is desire/value-dependent.

5.2.3 DV/SU *qua* Mb(DV)A and the Value of Utility

I began the last section with a brief reference to utility, its connection to preference, and the difference between imputing it along causal connections and symbolic connections. In this section, I want to discuss how a DV/SU account is utility-specific and how utility can be justified. Particularly, I want to argue for how one can justify the value in seeking utility, both the kinds of utility that are about outcomes and those that are bestowed on actions.

One argument that I made throughout the last section is that a value-based account or a value-sensitive view provides a better insight into the reasons for acting than either naïve or sophisticated desire-based accounts of reasons for acting. The upshot of this argument and my general analysis of desired-based and value-based views was that because DV/SU appeals to value, it better explains practical reasons than the theory of rational choice, of which Mb(CM)A is a special case. Although I think this analysis of desired-based and value-based accounts is adequate for my purpose of showing (in the next section) that DV/SU *qua* Mb(DV)A can dissolve the problem of secession, I believe it is fruitful to explore what makes utility worthwhile. To this extent, the reader must bear in mind that much of what follows in the remainder of this section, namely, my justification of what is valuable about utility is tentative, if not controversial.

Just to recap, SU refers to the utility an act (Φ) has intrinsically for its own sake, i.e. the expressiveness of Φ, and EU refers to the utility that Φ has for its instrumental value, i.e. the consequence of Φ. Since I am of the view that the DV account that parses utility into two components is accurate, the definition of utility that I have been working with so far requires some modification. I suggest a modified definition of utility that is action-sensitive yet reflects a consequence-sensitive definition of utility. Let us call utility on this modified view "the measure of *act-values*," that is to say, our relative preference or aversion for certain acts that have values for us. On this count, utility is still context sensitive and thus relative to our preferences.

If utility grounds preferences and is significant in cashing out reasons for acting, what is it that makes it valuable? To illustrate, suppose we point to the need to maintain and promote a certain sort of image or identify as the reason behind our choice of grooming, then it would be the case that the act of grooming in particular ways has some utility for us, otherwise why would it matter to us that we are groomed in one way rather than another.[344] In this case, the act of grooming in a particular way, we might say, is valuable because of what it expresses, in addition to the possible outcomes of the act. What the act expresses is positively intrinsically valenced for us and might be instrumentally valenced as well. But why is the utility of value?

There are several ways of responding to the demand that we demonstrate what is valuable about utility. First, we might simply say that the demand is wrongheaded and that what makes utility valuable is that it is *desired* by us and that there is no value outside the fact of our desire. The utility is o*u*r utility and it is of value because *we desire* it. Another way of responding to the demand, which is related to the first, is to say that what is valuable about utility is that it is *valuable for its own sake*. On this view, which may or may not draw on the rationality of telic occurrences, there is nothing valuable beyond the utility that connects to outcomes or actions. I choose to perform Φ because the utility it has for me is valuable for its own sake.

In general, we might say utility is valuable because of its congeniality to social life and practices, just in case congeniality to social life and practices is

[344] Sporting a Mohawk, or styling our hair in specific ways, wearing shaggy pants, or pants with certain designs would be examples of grooming in particular ways that we might say represent something about the groomer.

subjectively determined by us. There are two related ways we might formulate this. First, we might formulate the value of utility in terms of what it enables us to accomplish, where what we accomplish is anything that is related to what we enjoy doing. Second, the value of utility might be formulated as motivationally good, where the acts we perform require that we be so motivated. On the first formulation, utility only strengthens what we enjoy doing and on the second formulation, utility serves as an incentive, so to speak. Note that there is something eerily instrumental on both formulations. On both formulations, utility is valuable because it motivates us in the direction of beneficial social interactions. If utility is valuable in virtue of the fact that it encourages a vigorous pursuit of participatory social activities, then it would seem that the demand for the value of utility has been satisfied. This needs further explanation.

Suppose in a given society, burning the flag is negatively valenced and not burning the flag is positively valenced. Let us call the act of burning the flag, 'flag-burning act' and the act of not burning the flag, 'no-flag-burning act.' Suppose that 'no-flag-burning act' is positively valenced because it expresses certain beliefs or values about this society, then given that the act is positively intrinsically valenced, we would expect members of this society not to burn the flag. Because the act is intrinsically desired for what it expresses—and this expressiveness has utility—members in that society refrain from burning the flag. Suppose also that if this utility is nil, they will burn the flag. Suppose finally, that the utility in question is about the act alone and not about the outcomes of the act, namely, not about psychic thrills, good reputation, and community medals of good behavior or other

forms of instrumental reward. Because we have limited utility to the act and not to its outcomes, we are able to focus our attention on what the act expresses. In this society, one might assume that utility is valuable. When the utility is diminished or extinguished the act would not be performed and, consequently, the general practice of respect for the flag is effaced and annihilated.

Suppose an individual is an EU-seeker, that is, the individual is strictly motivated by EU-benefits, then we might assume that the individual would be averse to performing acts whose outcomes are believed to be nil. Similarity, suppose a person is an SU-seeker, that is, the individual is strictly motivated by SU-benefits, then we might assume that the individual would be averse to perform acts that are negatively intrinsically valenced. To put it in a different way, our SU-seeking interest—the driven tendency to engage in acts that express for us values we hold—forbids us from choosing those acts that do not express the values we hold, just as our EU-seeking interest—the driven tendency to engage in acts that maximize EU—forbids us from choosing those acts whose instrumental value are believed nil.[345] The application of DV to the PD in section 5.1 illustrates quite well this last point.

I believe we can apply the above line of reasoning to the flag-burning example. If members of the society in question have an SU-seeking interest, and if this interest points to, say, the value of respecting the flag, this will forbid them from burning the flag. And conversely, if they have no such SU-seeking interest or

[345] An incomplete account of practical rationality claims that we are one or the other, i.e. we are either EU-seekers (our interests are strictly EU-driven) or we are SU-seekers (our interests are strictly SU-driven). But a complete account—such as a DV/SU account of practical rationality—claims that we are both EU and SU seekers. Both sets of interests factor into our reasons for acting and what actions we chose significantly depends on the weights we attach to either sets of interests.

if their SU-seeking interest points to a different direction other than respecting the flag, they will go ahead and burn the flag. I can now make the following general point about utility. Utility from performing an act is of value if it is positively intrinsically or instrumentally valenced. It is positively intrinsically or instrumentally valenced if it moves us to behave in ways that are consistent with our social environment, specifically, if utility motivates us to behave congenially with our beliefs about social life and practices.

On this understanding, utility that is attached to the 'no-flag burning act' is valuable if we are sufficiently moved to abstain from setting the flag on fire, just in case the act is congenial to our beliefs about social life and practices. And utility that is about other acts, say, the act of lying or the act of being mean to other people, is valuable if we are moved to abstain from these acts, so long as these acts are congenial to our belief about social life and practices. Similarly, utility that is about the act of supporting those that are economically unproductive or the act of cooperating with less well-off members is of value if such utility moves the better-off members to interact with the less well-off members, just in case they believe such act is congenial to social life and practices. It is important to note that although congeniality to social life and practices gives us the basis of imputing value to utility, this in no way suggests that utility is objective. Utility is still subjective. It is subjective insofar as congeniality to social life and practices is subjectively determined by us.

5.3 Mb(DV)A's Multi-tracked Framework for Solutions

I am now going to develop fully the point I made in the last paragraph, namely, the point about the value of utility that attaches to the act of supporting less well-off members. I am going to do this by applying the characterization of practical rationality that I have being discussing so far to the test of application. That is, I shall be demonstrating how a desire-and value-dependent account of practical reasons like DV/SU *qua* Mb(DV)A provides a multi-track framework for solutions to the problem of secession.

In chapter four, I discussed the structure of the problem of secession and demonstrated Mb(CM)A's single-tracked silver bullet solution to it. I indicated that Mb(CM)A's foundation in rational choice theory makes it an attractive and perceptive contribution to the social contract tradition. This foundation, which comes with a stripped down description of expected utility, constitutes, as I noted, a narrow and misleading characterization of practical rationality. This characterization fails when applied to the problem of secession, suggesting thus the need for a revision of Mb(CM)A along an Mb(DV)A direction.

In the section that follows, I shall be subjecting Mb(DV)A's multi-track framework of solutions to the problem of secession. Immediately following my discussion in this section, I shall briefly make the case (in section 5.3.2) that affective morality and Mb(CM)A occupy different ends of the spectrum of silver-bullet accounts. In the final section, I conclude my discussion of Mb(DV)A and the problem of secession with a beautiful poem, "Table Talk," by Wallace Stevens, as

a way of helping the reader to ponder over the multi-tracked framework for solutions that Mb(DV)A offers in the test of application.

5.3.1 Mb(DV)A Multi-tracked Framework for Solutions and the Problem of Secession

That a DV is a multi-tracked framework for solutions was egregiously evident in our discussion in section 5.1 of Newcomb's Problem and the PD. As we saw with Newcomb's Problem, the specification of the relevant weights to each solution determines whether decision-makers choose one box (opaque box) or both boxes (opaque and transparent boxes). Also, with the PD, it was evident that when utilities other than those that appeal to EU are factored into the payoff matrix, decision-makers shift towards the cooperative solution rather than the dominance solution, as the change from figure 5.1b to 5.1f illustrates. In figure 5.1b, with only EU on the table, decision-makers were disposed towards the dominance solution. However, decision-makers were disposed towards the cooperative solution when their considered preferences or aversions for the acts are included into the payoff matrix, as figure 5.1f shows. They choose the cooperative solution rather than the dominance solution only because cooperation means or symbolizes some value for them.

There are a number of ways to characterize symbolic meanings, which I hereafter will call an SU event. We can say, as Nozick does, that "the action (or one of its outcomes) symbolizes a certain situation, and the utility of this symbolized situation is imputed back, through the symbolic connection, to the

action itself."[346] Kant holds that in acting morally one acts as a lawmaking member of the kingdom of ends. We can say that a particular action, say giving the correct change to one's customer symbolizes a certain moral situation—being a lawmaking member of the kingdom of ends—and this situation has some utility which flows back to the action. The difference between this sort of imputing back and that of rational choice theory is that the former's utility is imputed along symbolic connections, whereas the latter's utility is imputed along probabilistic or causal connections.[347]

Another way of characterizing the SU event is to say that the symbolic connection of an action to a situation enables the action to be expressive of some belief, attitude, or value. In the PD matrix, acting on the optimal solution by doing what is best for both decision-makers collectively expresses what we might broadly refer to as a cooperative attitude. In which case, what flows back is not 'raw utility' but expressiveness, namely, the expressing of some particular belief, value or attitude. Expressing this something has high utility for us, hence, we perform the symbolic action. A DV/SU *qua* Mb(DV)A account of practical rationality does employ both characterizations and we expect a broader or adequate decision theory to incorporate symbolic connections and meanings.

It is important to keep in mind the point about SU being act-sensitive. Being act-sensitive means that SU is the utility that arises from performing certain acts; it is the utility that an agent gets in virtue of performing an act (the act having

[346] Ibid, p.27.
[347] In economics and the other social sciences, because practical rationality is often defined in terms of desires, aims or preferences, the imputing of utility usually flows along probabilistic or causal connections.

a certain expressiveness) in addition to the possible outcomes of the act. There is a connection between choosing a particular action because of its symbolic value and choosing a particular action because it expresses a principle. What I mean by this is that the expressiveness or symbolization of an action can be such that choosing a particular action stands for all the other actions that the principle specifies. I might choose not to burn the flag because the action falls into the class of similar actions that express 'good citizenship' or 'loyalty,' for me, in which case my not burning the flag has symbolic value and utility for me.

Note, however, that the "standing for" of an action for a principle may go beyond an action representing other things of the same type of actions (other actions) or for a whole group of actions. My not buying certain clothes, say, because they are produced by child labor sweatshops may stand for all other actions or group of actions that represent 'injustice.' In addition to this, the action may also connect to other things that are not themselves actions, for instance, with being a certain sort of person or expressing one's belief about justice. In which case, doing the act now might affect the prospect of my repeating it or my estimate of the probability of doing it in the future.[348]

Mb(DV)A multi-tracked framework for solutions takes value and symbolic expressiveness and utility to be subjective and in this sense it is sensitive to context. For Mb(DV)A, a decision-maker's considered preference or aversion for an act is not independent of the utility that the decision-maker derives from performing that act. The utility from performing the act is SU. The utility is subjective because it is the utility that an action may have intrinsically or for its

[348] Nozick, *The Nature of Rationality,* p.26.

own sake. The subjective nature of value and utility was evident in the application of both CEU/EEU and SU to Newcomb's Problem and the PD. A decision-maker's choice of a solution in both problems was dependent on the weight or value that the decision-maker attached to the solutions: is there value in taking both boxes since there is only a Loonie in the transparent box? Is it worthwhile to rat on the other person when one gets only a marginal reward for one's behavior? Does the act of cooperation or the act of ratting symbolize any value, for one?

Mb(DV)A provides a multi-tracked framework for solutions to the problem of secession because it specifies not just the possible outcomes from secession acts and non-secession acts, but also factors in the considered preferences or aversions of well-off members for those acts. Now, Mb(DV)A does not claim that it is rational or not rational for productive members (citizens of the South) to cooperate with citizens of the North. Rather, it claims that what solution they choose (cooperation or secession) would be determined by their considered preferences or aversions for either solution. *If* the act of cooperation positively expresses *something* (values, beliefs about something) for them, taking into account the expected utility of the act, they will cooperate with citizens of the North; otherwise, they will secede.

An adequate or comprehensive account of practical rationality considers how decision-makers reason in choice situations and contexts, and models its conception of rationality according to this reasoning. Given that decision-makers do not reason only about the outcomes that are produced by the acts in choice situations; they reason as well about the values and meaning of the available acts, it

will be appropriate to consider what the acts of secession and cooperation means for citizens of the South, in addition to the outcomes of those acts. Mb(DV)A does just this; it takes into account how decision-makers reason in choice contexts when it claims that citizens of the South (better-off members) might have non-EU reasons, i.e. SU reasons for supporting those from the North (less well-off members), even though they might not have EU-reasons for cooperating with them. These reasons make it rational or not for them to support less well-off members.

The reasoning in terms of non-EU reasons for those from the South is explained by their considered preferences or aversions for the acts of secession and cooperation. The act of cooperation or secession is positively intrinsically valenced for citizens of the South, just in case the act has positive intrinsic utility for them, and the act has positive intrinsic utility for them if the valence of the act is independent of their belief that it brings about further outcomes that they desire. If the act of secession has no positive intrinsic utility for well-off members, the tendency to secede is extinguished.

To say that Mb(DV)A offers a multi-tracked framework for solutions and not a single-tracked silver bullet solution to the problem of secession is to say that it takes into account the full range of reasons that are relevant, from the point of view of decision-makers, in situations where secession looms large. If the act of cooperation or non-secession is positively intrinsically valenced for productive members, the tendency to secede is extinguished, but if cooperation is negatively intrinsically valenced, the tendency to secede is not extinguished. As was the case in Newcomb's Problem and in the PD, better-off members switch between both

solutions (cooperation and secession) depending on the value they assign to each solution. Note, however, that although Mb(DV)A offers a multi-tracked framework for solutions and not a single-tracked silver-bullet solution in the test of application, the framework separates situations in which secession is DV-irrational from those in which it is DV-rational (or may be so, when symbolic utilities are fully appealed to).

The picture that seems to emerge from an Mb(DV)A multi-tracked framework for solutions is that symbolic expressiveness and utilities are contingent. What I mean is this. That a person bestows *this* and not *that* symbolic expressiveness and utilities or *this* value and not *that* value to certain acts seems to be purely a contingent matter. There is a sense in which this is true. If my analysis of desire-based and value-based accounts in the preceding sections, and the argument in section 5.2.2 for desire-and value-dependent account of practical reasons—which connects reasons for acting with values (or symbolic expressiveness and utilities)—are correct, then the view that whatever values an individual bestows on acts are contingent seems about right. This is because an agent who takes his or her reasons for acting from values or an agent whose reasons for acting is sensitive to values has to subjectively determine those values if they are to motivate that agent to action. The corollary of this view is that if a person does not assign any value or symbolic expressiveness and utilities to the acts that are available, then what counts as decisive reasons for acting for that person will more than likely be taken from the possible outcomes of those acts. Hence, values or symbolic expressiveness and utilities count as reasons for acting

for an individual if and only if that individual bestows such values on acts or attaches symbolic expressiveness and utilities to such acts.

Although the kind of symbolic expressiveness and utilities to which a person appeals or the nature of the value that an individual assigns to acts are contingent in the sense described above, this does not mean that symbolic expressiveness and utilities *simpliciter* are not essential to who we are *qua* rational agents, or that symbolic expressiveness and utilities are the sort of things that a person might choose to have or not have. We may choose to associate with certain types of symbolic expressiveness and utilities and not others, and we may prefer this or that type of symbolic expressiveness and utilities, but we may not choose to *not* associate with any type of symbolic expressiveness and utilities. To say this is to say that we may not choose not to have any preferences. I might choose to have a Hemiwalker cane instead of a Quad cane because of my need for more support; however, Hemiwalker is not part of me, or essential to who I am. If by some improved medical practices I am restored to better health or walking condition so that I no longer have need to walk with a support, I am more likely to dispose for good with the use of any cane. Symbolic expressiveness and utilities are not like canes that can be disposed of for good.

If we think of SU as explained by utility as a special case, then SU is related to preferences in virtue of the fact that it is utility that explains or measures preferences. On this interpretation, SU as explained by preferences is essential to who we are and shaped both by our cultural environment and by our nature. It is true that in a sense what sorts of preferences people have are contingent, i.e.

contingent on their cultural environment and their nature. A person's preferences might be different if that person had a different body or lived in a different cultural environment. Similarly, if a person had a different body or lived in a different cultural environment, the symbolic expressiveness and utilities that a person has or associates with might be different from those that the person now has or associates with. However, the fact that our bodies and cultural environments give rise to whatever preferences we end up having does not make preferences contingent, as it is with symbolic expressiveness and utilities. Preferences are constitutive of rational agency and they point to reason-giving facts.[349] Preferences revealed by behavioral dimensions or choices and attitudinal dimensions or speech[350] help us to understand the human capacity for belief-and desire-representations. The capacity to represent beliefs and desires is theoretically fundamental to rational agency. As rational agents, we deliberate about possible actions in the light of our represented beliefs, desires or reason-providing facts and act in accordance with those deliberations. Because preferences help us to understand the human capacity for belief-and desire-representations, it is hard to imagine any rational being without such representations.

Social scientists and anthropologists, as Nozick rightly points out, have paid the most attention to the symbolic meanings of actions, rituals, cultural forms and practices as well as their importance within the ongoing life of a group.[351] They have provided us much to think about concerning how symbolic meanings

[349] Preferences or symbolic meanings are associated with a cultural environment when the meanings are those that the culture attributes to things, and they are associated with an individual's nature when they are the ones we ourselves bestow.
[350] See MbA, pp.27-29.
[351] Nozick, *The Nature of Rationality,* p.32.

configure and permeate our world. Living in such a rich symbolic or meaning-filled world, we draw on such meanings when we act. By our actions, we either reveal and expand or reject and escape these meanings. In conjunction with the meanings our actions have, we expand or escape the limits of our situations.

For an Israelite in biblical times, washing the feet of a guest or stranger *means* or *symbolizes* hospitality. As an Israelite, one imputes to the act of feet-washing utilities coordinate with what the act symbolizes. An Israelite may choose to perform the act of feet-washing or not. By performing the act, an Israelite strives to realize utilities and symbolic meanings associated with the act, and by refusing to wash a guest's feet, an Israelite either avoids or fails to realize these utilities and symbolic meanings. There are similar accounts of these sorts of practices in many cultures which connect acts with symbolic expressiveness and utilities. A familiar account which I discussed in chapter three is the practice of taking off one's shoes or leaving them at the door while visiting, a practice that many cultures take to express politeness or good manners.

Although the argument that values *qua* symbolic meanings permeate our world and that we draw on them when we act, as well as the related argument that those meanings are implicated in both our cultural environment and nature together explain the source of symbolic meanings and their expressions, these arguments do not explain how symbolic expressiveness and utilities are culture-spanning. Do members of a social milieu or cultural environment bestow the *same* meaning upon the same act? If symbolic meanings and their expressions are subjective (either bestowed by our culture or the ones we ourselves bestow), it is reasonable to ask if

everyone in that society confers the *same* meaning to the same act. Staring at someone in the eyes when talking to them might express politeness or good manners for an individual in a certain context, but does this act mean the same thing for everyone (in the same or in different contexts)? Since symbolic expressiveness and utilities are sensitive to context, would it not be right to say, for example, that feet-washing is not expressive of hospitality for *all* Israelites?

If we accept the view that people of the same cultural environment may not bestow the *same* meaning upon the same act, then the implication for the view that in situations where secession looms large, better-off members would choose to support less well-off members, provided that the meaning they assign to the cooperative act points toward that direction would be profound and telling. The implication would be that there is no convergence in the meaning of secession acts and cooperative acts. In which case, the argument that better-off members would choose to interact with and support less well off members is deceptive, if not utterly mistaken. For, what is clear is that some better-off members may choose to support less well-off members because the meaning they bestow upon the cooperative act points toward that direction, whereas other better-off members may choose not to interact with less well-off members because the meaning they assign to the cooperative act does not point towards that direction, but points toward the direction of secession.

Certainly, if what social scientists and anthropologists have told us about the importance of symbolic meanings of actions, rituals, cultural forms and practices to the continuing life of a group provides us with a rich nuanced approach

to social relationships, then we should pay attention to how such culture-spanning symbolic considerations affect where people gravitate, not just on issues of secession, but as well on a host of other fundamental social issues. In particular, there is no reason to believe that the history of interwoven relationships and interactions among citizens of the North and South, considered in the light of previous mutual interactions and bonds of interdependence of utilities and free affectivity, would not provide them with the sorts of common meanings or values for certain basic and important cultural forms and acts. The culture-spanning symbolic considerations may point in the direction of non-secession, for the North and South, even though EU may point towards secession for one group and non-secession for the other.

DV/SU *qua* Mb(DV)A enriches our understanding of rational agency and practical reasons. It helps us to understand why an individual might prefer or be averse to certain acts in addition to the possible outcomes of those acts. Intrinsic preferences or aversions about acts are not adequately explained by the possible outcomes of those acts. One may still stop by to help a stranger who has been mugged, even though one believes that there are no EU-benefits for that person in doing so—one might be late for a job interview or miss out on a once-in-a-lifetime music concert. One may refrain from burning the flag, notwithstanding the consequences of doing so. Burning the flag need not merely be instrumentally valuable, but rather it can be intrinsically valenced. Flag burning as an act would not be extinguished (pardon the pun) if its instrumental value were believed nil. One might still be averse to burning the flag regardless of its EU, just because the

act stands for something about oneself; if you like, it represents a certain value or principle about oneself, say, the value of 'good citizenship.'

Also, an individual may be prone to tell the truth, irrespective of the possible outcome of telling the truth. This idea about telling the truth independent of its possible consequences has been forcefully argued for by Kant in his moral ontology according to which the moral law that governs the realm of our moral lives ought not to be based on heteronomous grounds. Given thus, the objective and universal determination of the moral law, telling the truth for Kant, requires absolutely that a person tells the truth at all times. Hence, an individual is required, for example, to tell the 'inquiring murderer' where his victim is headed, notwithstanding the consequences of the action, just in case the maxim of that person's action satisfies the requirement of the categorical imperative. DV/SU agrees with Kant that one may tell the truth in spite of the possible outcome of telling the truth. However, DV/SU does not require absolutely that we tell the truth. Since DV/SU locates our reasons for acting on the values or symbolic expressiveness and utilities of those acts, whether we tell the truth or not would depend on what the act of telling the truth means for us. On this view, telling the truth may mean telling the 'inquiring murder' where his victim is headed, regardless of the possible outcomes of the act, just in case the act expresses something for the individual and what this act expresses has utility for him or her.

To illustrate the point about instrumental and positive intrinsic utilities and how they are implicated in the reasoning of decision-makers in choice contexts and in normal social practices, consider the argument often advanced in support of the

intrinsic value conception of democracy. The claim that democracy is intrinsically valuable is often defended because the instrumental value or 'service conception' notion of democracy is considered inadequate. According to the instrumental value conception of democracy, democratic processes and institutions are valuable or adopted because they lead to good outcomes. The common objection is that the instrumental value justification of democracy does not discriminate between sound political and policy decisions, i.e. good outcomes reached democratically, and perfectly identical (and identically sound) decisions reached, say, by the arbitrary will of some authoritarian ruler. The intrinsic value conception claims that democratic processes and institutions are valuable not simply because they lead to good outcomes but because they embody the value of equality and freedom.

As a case study, consider the South African election of 1994 that brought Nelson Mandela to power. This election was celebrated worldwide. Black South Africans were enfranchised for the very first time. But why was this viewed as significant? Without doubt, this was not because of the outcome of the vote; it is not simply that under the new political environment, black South Africans were more likely to be treated with justice than they had been in the past, although this was surely part of the reason for the worldwide celebration. Rather, it seems that the main reason for celebration was that including black South Africans in the electoral or democratic process demonstrated that black South Africans were finally being treated as equals and as worthy of respect. That people are included in the democratic process has a certain intrinsic or expressive value. In this case, it symbolizes that in some way at least black and white South Africans stand together

as social and political equals. Having a vote, then, seems to be important irrespective of what outcomes the vote produces.

From the forgoing, we can isolate at least two conditions that have to be satisfied for a symbolic action to be done. First, the action in symbolizing a situation must provide some utility to the individual who chooses the action, in other words, utility has to flow back along symbolic connections or lines for that individual. This symbolizing takes into account the claim that reasons for acting are to be given by values or facts that make certain possible outcomes worth producing or preventing. The utility that flows back along symbolic connections constitutes the reason-providing fact for the action. In figures 5.1e and 5.1f, utility flows back along symbolic connections once we take into account decision-makers' considered preferences and aversions for the available acts. Second, the utilities that an individual derives from choosing a particular act have to be greater than the utilities of other available acts. If Φ_1 has **7 EU** and Φ_2 has **9 EU**, and if both acts have no SU, an EU-focused account stipulates we take Φ_2, and DV generally agrees. But if Φ_1 has **5 SU** in addition to the **9 EU**, then on the strength of the higher utilities of Φ_1 (**7 EU** + **5 SU**), DV/SU stipulates we choose Φ_1.

This does not mean that symbolic expressiveness and utility are subordinate to EU or lexically preceded by outcome-utilities. An action might still be intrinsically valenced and hence performed even when its instrumental value is nil. As the flag-burning example shows, one would still be averse to burning the flag despite the content of the causally-produced consequences. Sometimes symbolic meaning might even be thought better than causal consequences. If an outcome

such as harming someone in revenge (to use Nozick's example) is "desired but seen as bad, it may be better for a person to achieve this symbolically than to inflict actual danger."[352] At other times, symbolic meanings might serve as a tiebreaker in cases of indeterminacy like that of the Buridan ass or in some other paradox-like situations. Suppose one is immortal and has in possession a bottle of 'EverBetterWine.' The wine improves with age. In fact, it improves so steadily and so rapidly that no matter how long one waits before drinking it, one would be better off, all things considered, waiting one more day.[353] In this situation, symbolic meaning would crucially determine when to drink the wine. Were one to ignore symbolic meaning one would have no reason on an EU-focused account to drink the wine since every single day one waits, one improves one's EU profile. But if one considers symbolic expressiveness—it is not nice to have your guests over and not serve them wine (especially when your guests have just had chicken and the white wine is sitting on the shelf)—then choosing to drink the wine on the day your guests had just had chicken seems reasonable.[354]

SU shines through more in cases where EU is stacked too high against performing certain actions. The cases I have in mind are situations where people prefer an action or persist in choosing an action in the face of strong evidence that the action has negative causal outcomes, or does not actually have the presumed causal consequences or its outcomes are zilch. The dire consequences people bear

[352] Ibid, p.31.
[353] This example is taken from John Pollock, "A Theory of Moral Reasoning," in *Ethics* (1986), pp.506-523: 517.
[354] If this way of putting the example is correct, symbolic meaning and utility may provide us a way of incorporating satisficing within a maximizing conception of practical rationality. I leave this for another occasion.

in order to avoid 'losing face'; self sacrifice in war despite negative consequences; the death people—samurai warriors, gladiators—risk to 'maintain honor'; or, perhaps even Socrates' refusal to avoid the penalty of fleeing to Thessaly—all seem to fit the cases of SU shining through where EU is stacked too high against performing certain actions.

To take up the example of Socrates, we could ask the question, does his refusal to avoid the penalty of fleeing to Thessaly and his drinking of the hemlock have any EU weight? Within the broader context of an EU-focused account, Socrates' refusal to avoid the penalty of fleeing to Thessaly is irrational. This is because there is not much EU in drinking the hemlock, but there is much EU in being alive in Thessaly. Indeed, EU is stacked quite high against Socrates drinking the hemlock. So Socrates is EU-rational if he flees to Thessaly, and EU-irrational when he drinks the hemlock. However, within the broader context of decision-value, DV/SU says that Socrates' refusal to avoid the penalty of fleeing to Thessaly and his decision to drink the hemlock are rational and may be irrational depending on what the acts of running away or drinking the hemlock means to him. Hence, to properly understand his behavior we have to factor in his considered preference for the acts of fleeing to Thessaly and of drinking the hemlock. If one supposes that he drank the hemlock because the act expresses for him certain values—say, the value of 'nobility,' or of 'acting nobly,' or of being a 'good citizen,' or some other values—and this expression has utility for him, then it would be the case that although his behavior has no EU weight in the sense that it is EU-irrational, it has SU weight in the sense that it is DV-rational.

How does Mb(DV)A fare with the issue I raised in chapter one, namely, the issue of the tension between liberal individualism and communitarianism? How might we think of it as navigating a middle way between liberal individualism and communitarianism in general and between the strict individualism of Mb(CM)A and extreme communitarianism of Rousseau's social contract theory? Rousseau's extreme communitarianism identifies individual interests with collective interests. The *citoyen* sets aside the egoism of the *primitive individual* when she makes the transition from *noble savage* to *citoyen*. In the transition to 'a new person,' she loses her individuality, which effectively becomes identical with the general will. Mb(CM)A's strict individualism identifies individual interests with the maximization of that individual's EU. The liberal individual, according to Mb(CM)A, has the affective capacity for (rational) morality. She makes the transition from the economic individual to the individual that she now is, but in this transformation, she does not lose her individuality.

The liberal individual of Mb(CM)A is neither a candidate for Rousseau's general will nor is she Hume's ideal sympathizer. Unlike the *citoyen* and the ideal sympathizer, the liberal individual instrumentally values fellow participants and participatory social activities. She possesses a sense of duty and she values other participants and participatory social activities because of the value she places on the benefits these provide her. Because she values other participants and participatory social activities, the liberal individual recognizes the importance of mutually beneficial activities and thereby develops and manifests tuistic feelings and bonds necessary for such activities. Unlike the liberal individual who sees duty

and justice as a means of promoting her rational self-interests insofar as this is compatible with the rational self-interests of others, duty and justice for the *citoyen* means identifying her interests with general interests.

Mb(DV)A navigates a middle way between Rousseau's account and Mb(CM)A by not identifying an individual's interests with the maximization of that individual's EU, but with the maximization of DV, such that the individual values participants and participatory social activities not just for the instrumental value but as well for the intrinsic value that these hold for her. That is, she values participants and participatory social activities not just instrumentally, but intrinsically for their own sake. Mb(DV)A acknowledges the significance of the transition from the economic individual to the liberal individual, but holds that the liberal individual, as defined in Mb(CM)A, has not been adequately described. In order for her to relate suitably to other participants and to exploit participatory social activities for mutual benefits, Mb(DV)A proposes a transition from the liberal individual in (Mb(CM)A to *a communally-tempered liberal individual*.

In navigating a middle way, Mb(DV)A, places the interests of the individual not 'outside' the interests of others, but 'within' their interests. The *communally-tempered liberal individual* recognizes the importance of general interests (captured in Rousseau's general will) and individual interests (captured in Mb(CM)A). She recognizes general and individual interests when she seeks to maximize her interests within a framework of value that speaks both to outcomes and her considered preference for the acts that produce those outcomes.

To say that the *communally-tempered liberal individual* of Mb(DV)A embraces participants and participatory social activities not strictly for their instrumental value, but also for their intrinsic value is to say that the *communally-tempered liberal individual* is disposed both *individually* and *communally*: disposed individually when the acts she chooses are those that symbolize something for her, i.e. provide utilities for her, and disposed communally when the acts she chooses are those that provide utilities both to her and others. Specifically, the *communally-tempered liberal individual* associates her interests with others when doing so maximizes DV for her, and she recognizes that in maximizing DV she benefits herself at the very same moment that she benefits others. By maximizing DV, the *communally-tempered liberal individual*, *simpliciter*, maximizes both her interests and the interest of others.

My argument up to this stage points in one direction, namely, that an Mb(DV)A multi-tracked framework for solutions to the problem of secession replaces the Mb(CM)A EU-focused single-tracked silver-bullet solution to the same problem. The crucial aspect of Mb(DV)A's replacement of Mb(CM)A is located in the claim of what it is that we maximize as rational agents when we act. For Mb(CM)A, what we maximize when we act is EU, but for Mb(DV)A, what we maximize is DV. We maximize EU when we *only* take into account the possible outcomes of the available acts, but we maximize DV when in addition to the possible outcomes of the acts we take into account our considered preference or aversion for those acts.

Hence, it may or may not be in the interest of better-off members to interact with or support less well-off members, but whether better-off members choose to cooperate with less well-off members or whether they consider it in their interests to support citizens of the North would depend on what the cooperative and secession acts mean for them. When EU is stacked too high against cooperation, it is not rational nor in the interest of better-off members to cooperate. However, taking into account their considered preferences or aversions for the cooperative and secession acts, it is rational or not rational for them to interact with and support less well-off members depending on whether they prefer or are averse to cooperative or secession acts or depending on which way the SU of each act points toward. The bottom line of Mb(DV)A's claim is that SU may, in general, unite people with respect to certain reasons, even if we suppose that they may be divided in, say, their EU-orientation. For if the meaning that better-off members bestow upon cooperative acts points toward cooperation, then they will choose to interact with and support less well-off members, and if the meaning they assign to cooperation points toward secession and not cooperation, then they will choose not to interact with less well-off members.

My argument for Mb(DV)A thus demonstrates that although I agree with the first part of Braybrooke's negative thesis, I do not accept the second part of his negative thesis, which is that any social contract theory cannot resolve the problem of secession. My argument for Mb(DV)A was in effect an argument for why Mb(CM)A breaks down in the test of application as well as an argument for how a modified moral contractarian account of reasons like Mb(DV)A resolves the

problem of secession. When we interpret Mb(CM)A along Mb(DV)A, serious objections such as the problem of secession fall away. Mb(CM)A is grounded on EU-reasons, such that when expected utilities are stacked too high against cooperation, it is not rational to cooperate. Mb(DV)A revises this narrow view of practical reasons by factoring symbolic utility and expected utility into the reasons for acting, such that when expected utilities are stacked too high against cooperation, it may or may not be rational to cooperate depending on the direction that symbolic utility points toward.

5.3.2 Affective Morality and Mb(CM)A as Silver-Bullet Accounts

I can now venture into making the claim that affective moralities and Mb(CM)A are accounts that occupy different ends of the spectrum of silver-bullet accounts. In making this claim, my discussion touches tangentially on Braybrooke's claim regarding affective moralities and the problem of secession. In chapter four, I discussed what Braybrooke's negative thesis is: (a) Mb(CM)A cannot resolve the problem of secession (b) any social contract theory cannot resolve the problem of secession. In arguing for Mb(DV)A, I argued that negative thesis (a) is true, but not negative thesis (b). Braybrooke's positive thesis is that a theory of moral sentiments can resolve the problem of secession. I reject the positive thesis. I provide reason for this in what follows.

In rejecting Braybrooke's positive thesis, I claim that a theory of moral sentiments of the type defended by Hume (to which Braybrooke appeals) is similar to Mb(CM)A when applied to the problem of secession. Both a theory of moral

sentiments and Mb(CM)A, which occupy different ends of the spectrum of silver-bullet accounts, offer a single-tracked silver bullet solution in the test of application. Both appeal to a different framework, shall we say a single 'principle' or 'value.' And the principle to which each account appeals moves them to different solutions. On the one hand, a theory of moral sentiments appeals to sympathy or needs in order to resolve the problem of secession; sympathy or needs drive people to choose non-secession acts. On the other hand, Mb(CM)A appeals to EU to resolve the problem of secession; EU drives people to choose secession acts. On both accounts, they track a single 'principle' and this leads them to recommend different acts: secession for Mb(CM)A and non-secession for affective morality. However, on these accounts the decision-maker's preference or aversion for those acts that are available is egregiously not accounted for in.

In chapter four, I examined Hume's theory of moral sentiments as a way of demonstrating how a moral account built upon 'fixed affectivities' is different from Mb(CM)A, which is grounded in free affectivities. A theory of moral sentiments claims that given that better-off members are ideal sympathizers, they will in situations of secession support less well-off members. Because better-off members possess sympathetic feelings that have been properly calibrated to recognize 'situations of need,' they will extend their agreement widely to include less well-off members. Or, as Braybrooke pointedly puts it, contractors—both the well-off citizens of the South and the less well-off citizens of the North—will accept the terms of the contract not on the basis of preference-satisfaction but on the basis of needs-satisfaction. He boldly

proclaims: we "heed anybody's preferences only insofar as it is consistent with meeting everybody's needs."[355]

Braybrooke may choose to champion an affective needs-based social contract that restricts the scope of the considered coherent preferences of agents in favor of mutual concern for needs. He may choose to defend an unrefined conception of self-interest, one that incorporates sentiments. But by defending an unrefined conception of self-interest, he defines self-interest narrowly and in doing so he assumes away what human or rational agency essentially requires, namely, the capacities to act and make choices as one sees fit and to impose those choices on the world. By assuming away rational agency, Braybrooke presents us with a silver bullet account, one that fundamentally identifies with a single 'principle': needs-satisfaction.

Mb(CM)A distances itself from moralities of fixed affectivities, and rightly so. In defending the essentially just society as the society made up of liberal individuals, Gauthier argues that moralities of free affectivities encourage cooperation in ways that are different from affective moralities. Whereas moralities of free affectivities encourage cooperation among members of society as long as such cooperation is mutually advantageous, affective moralities encourage cooperation not because cooperation is mutually beneficial but because of the possession of emotional responsibilities by members of society. Affective moralities command minimal assent because emotional responsibilities are thrust upon persons independently of their preferences and volitions. And as Gauthier correctly observes, this is destabilizing for society since emotional responsibilities

[355] Braybrooke, "Social Contract Theory's Fanciest Flight," *Ethics*, pp.762, 763.

are not voluntarily chosen. For, in the absence of freely chosen affectivities, people can hardly be called upon to honor agreements, not just when there are no constraints but also in situations where expected utility is stacked too high against cooperation.

Mb(DV)A agrees with Mb(CM)A that affectivities should be related to the individual, that is to say, affectivities should be determined by the individual if we are not to assume away what rational agency requires and if we are to preserve the thesis of individualism that is central to any liberal framework of social cooperation. However, Mb(DV)A distances itself from Mb(CM)A because of its narrow and misleading characterization of practical rationality. Affective moralities invoke needs-satisfaction or emotional responsibilities to argue for the reason for the support of less well-off members. Mb(CM)A does not do this. It invokes instead EU to argue for the reason for the non-support of less well-off members. Affective moralities and Mb(CM)A's morality of free affectivities have one common problem: both separate rationality from emotion. Mb(DV)A challenges this separation, doing so without abandoning the rigors of decision theory. In doing this, Mb(DV)A makes the case for *emotion* 'within' decision theory. Emotion within decision theory takes the targets of emotions as decisive not because of their outcomes but because of what they convey about the meaning of the emotional attitudes of the decision-maker.

In making the case for *emotion* within decision theory, Mb(DV)A claims, that reason could point towards both directions in situations where EU is stacked too high against cooperation, when we factor in the meaning of the emotional

attitudes of the decision-maker. Better-off members may choose not to support less well-off members depending on the value or meaning they bestow on secession and cooperative acts (factoring in as well the meaning of their emotional attitudes), and they may choose to cooperate with less well-off members depending on the value they bestow on secession and cooperative acts (factoring in as well the meaning of their emotional attitudes). And if we suppose that choosing to support less well-off members is positively intrinsically valenced for better-off members, then they would choose to cooperate, but if we suppose that choosing to support less well-off members is negatively intrinsically valenced for better-off members, then they would choose to secede.

5.3.3 'Table Talk' by Wallace Stevens[356]

Granted, we die for good.
Life then is largely a thing
Of happens to like, not should.

And that, too, granted, why
Do I happen to like red bush,
Gray grass and green-gray sky?

What else remains? But red,
Gray, green, why those of all?
That is not what I said:

[356] Wesley Copper suggested this as a possible section header.

Not those of all. But those.

One likes what one happens to like.

One likes the way red grows.

It cannot matter at all.

Happens to like is one

Of the ways things happen to fall.

So the citizens of North and South may like what they 'happen to like.' They may 'happen to like' each other. They have beliefs about value, such as 'there but for the grace of God go I,' that infuse those likings with value. These likings transcend EU-reasons and are not needs-based nor circumscribed by sympathetic feelings. They are circumscribed rather by values or what values express for them. I am proposing a reading of Mb(CM)A that shows how this situation can be rational. On the proposed reading, the situation involving the citizens of North and South and their likings need not be irrational sentimentality. I am replacing Mb(CM)A with Mb(DV)A, an account that offers a multi-tracked framework for solutions in the test of application to the problem of secession and shows how citizens of North and South may 'happen to like' each other in ways that are rational.

Chapter Six

Critiquing Mb(DV)A's Multi-tracked Framework for Solutions

Introduction

In this chapter, I will examine three critiques of the Mb(DV)A's multi-tracked framework for solutions I have defended in this work. These are: (i) how do we know which symbolic meanings and preferences are desirable or good, and which ones are undesirable or not good?; (ii) suppose a person can be caused to have various symbolic meanings and preferences, should that person be shaped to have SU-reasons for being united with others, or for cooperating with others?; (iii) does Mb(DV)A not violate the demand of an essentially just society as a cooperative venture for mutual advantage?

6.1 How do we Know Which Symbolic Meanings and Preferences are Good or Desirable?

Critique

According to Mb(DV)A, we draw on symbolic meanings and their expressions when we act. It claims that a large part of the richness of our lives consist in symbolic meanings, and their expression, namely, the symbolic meanings our culture attributes to things or the ones we ourselves bestow. Suppose we accept this view, and since not all symbolic meanings and preferences might be good, how do we know those that are good, that is to say, if we are to draw on good symbolic

meanings and their expressions, how do we distinguish between those that are good and desirable and those that are bad and undesirable?

Response

In his discussion of symbolic expressiveness and utility, and the nature of rationality, Nozick acknowledges that a theory of symbolic meanings and preferences is needed to separate desirable or good symbolic meanings and preferences from those that are undesirable or not good. He writes:

> Notice that symbolic meanings might not all be good ones, just like desires or preferences might not be. The point is that a theory of rationality need not *exclude* symbolic meanings. These do not guarantee good or desirable content, however. For, that, one would need to develop a theory of which symbolic meanings and which preferences and desires are admissible, using that to constrain which particular meanings and desires could be fed into the more formal theory of rationality.[357]

It is a legitimate critique that for an account of symbolic meanings and preferences to be theoretically compelling, we would need a framework for discriminating between good or desirable symbolic meanings and preferences and bad and undesirable symbolic meanings and preferences. This will be the direction of my future research. But for now, it is important to recognize that symbolic meanings and their expressions, as Nozick rightly points out, play a decisive role in the reasons for acting for agents, and as such a theory of rationality ought not exclude them.

[357] Nozick, *The Nature of Rationality*, fn p.30.

In this project, I did not set out to primary demonstrate what preferences and symbolic utilities are rational, desirable and good. What I set out to do was to show that Mb(DV)A provides a framework for solutions, when both symbolic utility and expected utility are sufficiently appealed to, in choice contexts in general, and in situations where issues of secession loom large, in particular. Specifically, the task I undertook was to demonstrate, in general, that Mb(DV)A provides a framework for discriminating between situations that are DV-rational and situations that are DV-irrational (when symbolic meanings and preferences are sufficiently appealed to), and to demonstrate, in particular, how cooperation can be DV-rational and DV-irrational (when symbolic meanings and preferences are sufficiently appealed to).

It is my expectation that future research into symbolic meanings and preferences will provide valuable insight into how to fashion a robust theory of symbolic meanings and preferences. And that the theory would provide a framework for a number of topics in rational choice theory and morality: topics like symbolic utilities and preference, and maximization/optimization and satisficing. Specifically, the theory would provide a framework for (1) distinguishing good or desirable symbolic meanings and preferences from bad and undesirable symbolic meanings and preferences and (2) for separating between situations where it is good and rational to satisfice and situations where it is not good and rational to satisfice, namely, situations where it is good and rational to maximize.[358]

[358] Broadly defined, a satisficing view claims that one is *permitted* to choose an action that implements a "satisfactory" or "good enough" means to one's given ends. A Maximizing view

6.2 Is it in a Person's Interest to be Shaped to have Symbolic Meanings and Preferences?

Critique

Suppose X has Y's best interests at heart and could shape Y's character, values and preferences. Perhaps X is a parent with great psychological insight, who could cause Y to have or not to have various values or symbolic meanings and preferences. Should X shape Y's character, values and preferences so that Y's considered preference and aversion for acts follow a pattern, or shape Y so that Y has the SU-reasons for being united with others, or for cooperating with others? What would be best for X and Y?[359]

Response

There are two issues here. First, should X shape Y to have particular values, preferences, or SU-reasons for being united with others, or for cooperating with others? Second, would it be best for X and Y if Y were shaped to have these preferences, values or symbolic meanings, or would it be best for both of them if X does not cause Y to have these preferences, values or symbolic meanings? Now considering the claim of Mb(DV)A that a large part of the richness of our lives consist in symbolic meanings, and their expression, it seems descriptively accurate to say that our socialization process aims towards shaping people to have particular values or symbolic meanings and preferences.

---

rejects this, and argues that one is *always required* to choose an action that implements the best means to one's given ends.

[359] This critique is suggested by Adam Morton.

For instance, our social upbringing seems to 'thrust upon us' various values or symbolic meanings, and their expressions that are often bestowed upon different things: practices, cultural forms, actions and rituals. And we are generally expected to implicate these values or symbolic meanings, and their expressions in these practices, cultural forms, actions and rituals. As, an example, consider again the Israelite in Biblical times who is brought up to recognize the value or symbolic meaning of washing the feet of a guest or stranger. Now, is it a good thing for a young Israelite to be so socialized, namely, should she be shaped in such a way that she comes to hold the value or symbolic meanings associated with washing the feet of a guest or stranger? Is it in the best interest of the young Israelite, the society and those who shaped her to have such value or symbolic meaning? It may or may not be beneficial for the young Israelite, her family and the society for her to be socialized to hold the value or symbolic meanings associated with washing the feet of a guest or stranger, or to be expected to implicate this value or symbolic meaning in various practices, cultural forms, actions and rituals.

In view of the fact that the primary aim of our socialization process is directed towards causing us to have various values or symbolic meanings and preferences, and given that we have to individually determine those values or symbolic meanings and preferences that we act on if they are to count as valuable, the question whether X should cause Y to have these values or symbolic meanings and preferences and not others, or to have any particular values or symbolic meanings cannot be answered in a straightforward manner. The answer to the question whether X ought to shape Y so that Y has the SU-reasons for being united

with others, or for cooperating with others ultimately depends on whether on reflection, X, as well as the society at large, considers it important for everyone in the society to act in ways that benefit the society, i.e. keep the society united, and if X believes that shaping Y in the way in question (i) contributes to this goal and (ii) benefits Y. Bear in mind that eventually, those who have been so shaped to have particular values or symbolic meanings and preferences would have to individually determine them as valuable when they act.

Moreover, given our discussion of the value of utility in section 5.2.1, i.e. that utility is of value to us just in case it motivates us to perform acts that are congenial to social practices, we might say that X ought to cause Y to have SU-reasons for being united with others, or for cooperating with others just in case cooperating or interacting with others is congenial to social life and practices. But beyond the comments about utility, value and the individual determination of utility and value we might say that the question whether X ought to cause Y to have particular values or symbolic meanings and preferences, and what sorts of values or symbolic meanings and preferences that X ought to cause Y to have would require a theory of symbolic meanings and preferences. If that theory is able to provide a framework for discriminating good or desirable values or symbolic meanings and preferences from bad and undesirable values or symbolic meanings and preferences, then we might be able to say what is good about this or that symbolic meanings and preferences and why one should be shaped to have them.

6.3 Does Mb(DV)A Violate the Demand of Mutual Advantage?

Critique

Mb(DV)A claims that when expected utilities are stacked too high against cooperation, and SU points toward cooperation, it is rational for better-off agents to interact with and support less well-off agents. Does this account not violate the demand of mutual advantage that is fundamental to contractarianism?

Response

Recall that for Gauthier a contractarian morality is a morality of mutual advantage. On this view, it is in the interest of better-off agents or not rational for them to interact with less well-off agents. In the North-South example, it is in the interest of those from the South to neither interact with nor support those from the North in their unproductiveness. On Mb(CM)A EU-focused account, the reason why it is not rational for citizens of the South to support citizens of the North is that such act does not maximize their expected utility. But this is not all. Providing support for the North fails the demand of an essentially just society, which requires that participatory social activities and schemes of cooperation be mutually advantageous.

Certainly, if the flames of morality are to be kept burning, those involved in keeping it aflame must benefit from the gains that morality promotes. This is the demand of mutual advantage, which is satisfied when everyone that contributes to participatory social activities or morality is benefited by such participation. If society as Rawls and Gauthier individually note, is a cooperative venture for

mutual advantage, then mutual morality is constitutive of society. If it is the case that extending our agreement more than benefits us encourages freeriding, and if this violates mutual morality how then can one justify the position that if secession acts are positively intrinsically valenced for better-off members, regardless of whether the instrumental value of those acts are believed nil, they would choose to interact with less well-off members? To state this in a slightly different way, how do we justify that the act of providing support for citizens of the North—assuming that such act expresses something of value for those from the South—does not violate the demand of an essentially just society as a cooperative venture for mutual advantage?

Certainly, extending support to unproductive and less well-off members appears inimical to the mutual morality that Mb(CM)A promotes, but does the fact that it is opposed to the morality of Mb(CM)A means that it fails the demand of an essentially just society as a cooperative venture for mutual advantage? I agree with Gauthier that mutual morality is constitutive of society. I also agree with him that a scheme of cooperation satisfies mutual morality when everyone that contributes to that scheme or participates in it benefits from such participation. However, I disagree with the view that extending support to less well-off members fails the demand of mutual advantage. It is unquestionably true that it is not in the EU-interest of better-off members to support less well-off members, but it does not follow from this that the act of providing support for citizens of the North violates the demand of mutual advantage.

In what follows, I am going to defend Mb(DV)A against the charge that it undermines mutual advantage. My aim here is to show that Mb(DV)A is consistent with the view that a scheme of cooperation satisfies mutual morality when everyone that contributes to that scheme or participates in it benefits from such participation. I shall frame my arguments for why I think that Mb(DV)A does not undermine mutual advantage accordingly. First, I will argue that mutual advantage does not require reciprocal performance of actions, that is to say, mutual advantage can sometimes be satisfied by performing unilateral actions. Second, I shall argue that mutual advantage can be understood in two senses: the *strong sense* and *weak sense,* and that Mb(DV)A satisfies the latter. Whereas, the strong sense appeals to expected utility, the weak sense appeals to symbolic utility. By making this distinction, I hope to be able to demonstrate that the weak sense of mutual advantage justifies the rationality of extending support to unproductive and less well-off members.

I begin by stating what I take to be a key difference between reciprocal performance of actions and unilateral performance of actions. In reciprocal performance of actions, two parties mutually benefit each other by 'doing something,' i.e. by each performing different actions. In unilateral performance of actions, one party benefits everyone by 'doing something,' i.e. by performing a specific action. There is no reason why mutual advantage cannot be extended to unilateral performance of actions, namely, to situations where one person performs an action that benefits others. But how can unilateral performance of actions be

mutual when only one party performs an action for the benefit of others? And how does the performance of unilateral actions discourage freeriding?

It would be apposite at this juncture to appeal to Rousseau's view of collective altruism. Bear in mind the distinction made in chapter one between 'collective altruists' and 'de facto altruists.' Individuals, according to Rousseau, are collectively altruistic when they identify their interests with their truly essential freedom by creating themselves 'anew' in the general will. And they create themselves as a 'new person' when they benefit others at the very same moment they benefit themselves. Without endorsing Rousseau's collective altruism in its entirety, we can embrace the idea that an individual can perform an action that simultaneously benefits that individual and others. Starting with this idea of an individual performing an action that at the same time benefits that individual and others we can extend the reasoning to unilateral relationships, where one person performs an action that benefits everyone even though the action holds out no possibility of reciprocity.

In any human interaction, if a person performs an act, say, the person mows someone else's lawn and the act provides the performer and recipient utilities, then such relationship can be considered mutual. Conversely, if the performed act provides the recipient and not the performer utilities, then there is a sense in which the relationship might be said not to be mutual. For when a person receives no utilities from performing an action, or when the act benefits the recipient and not the performer, the act and relationship cannot be said to be mutually advantageous.

But in what way can it be said that a unilateral performance of an action benefits the recipient at the very same moment the performer is benefited?

In principle, a person can benefit from performing an action that has no possibility of being reciprocated. There are many examples we can draw on to illustrate this. The targets of charitable acts are people in needs, but the performer of those acts can be said to benefit as well from performing the acts. Does Mother Theresa or the moral philanthropist not benefit from performing charitable acts? One might debate whether or not charitable acts fall into the rubric and category of altruistic or egoistic acts, but one cannot debate whether they fall under unilateral performance of actions. In the same line of thought, one might argue whether or not Mother Theresa or the moral philanthropist is altruistic or egoistic, but one cannot argue that they benefit from unilaterally performing charitable acts if it so happens that they derive pleasure from performing those acts. If I gain in one way or another from performing an act and if this act targets you, the mere fact that I gain or I receive benefit from performing the act gives us reason to claim that the interaction is mutually advantageous. One might even suggest that the benefits that I get from performing the act constitutes some sort of 'contribution' from you, for in your absence there would be interaction, and in the absence of interaction I would have no benefit.

Thinking of mutual advantage in the Mother Theresa or the moral philanthropist example involves cashing benefits in terms of happiness, pleasure or psychic thrill, which Mother Theresa or the moral philanthropist gets from unilaterally performing charitable acts. Note though that these are EU benefits. In

which case, we might suppose that the unilateral performance of charitable acts by Mother Theresa or the moral philanthropist satisfies mutual advantage. If mutual advantage is satisfied, it is satisfied because Mother Theresa or the moral philanthropist benefits those in need at the very same moment they benefit themselves. If this is right, then we might suppose that the crucial feature of mutual morality is not who performs what acts, or whether B reciprocates the action of A, but whether in performing specific acts everyone (A as well as B) is benefited. This way of putting it suggests that it matter not that you do nothing for me when I perform unilateral actions, for in performing those acts I get some benefits (and so did you), benefits that I otherwise would not have were I not to perform those acts.

We might even say that you would not have received those benefits if my situation has not called for you to perform those acts. In a world of plentitude or a world where everyone is self-sufficient, charitable acts may be unnecessary and otiose. In such a world, Mother Theresa or the moral philanthropist will not perform charitable acts. Having been deprived of the opportunity to perform charitable acts, they would miss out on the benefits that performance of those acts offer. I do not have a lawn to mow because I live in a condominium, and if not for my mowing your lawn, I would not have the benefits that I now have. I mow your lawn and I am benefited, and like Rousseau, we say, in benefiting myself, I benefit you at the very same moment. If we live in a society where they are no lawn or grass, then I would not have the benefits that mowing lawns offer.

But suppose I do not receive any EU-benefits (no happiness, pleasure or physic thrill) from performing the act, or my mowing your lawn gets in the way of

me being able to engage in some other acts that presumably provide me higher EU, that is acts that increase my happiness and utility profile. I might be happier or I might get more pleasure counting the blades of grass in the lawn. In the face of diminished or almost extinguished EU, that is to say, when EU is stacked too high against mowing your lawn or against cooperation, or performing certain acts it would seem irrational for me to mow your lawn, cooperate or perform those acts. This seems to be the basic argument of Mb(CM)A for mutual morality. If Mother Theresa or the moral philanthropist are rational, which we assume they are, and if they are EU-seekers—if you like, say, they want more pleasure or happiness than less—and if performing (unilateral) charitable acts diminishes or almost extinguishes their EU-seeking interests, or if they get greater EU by doing something else—such as playing golf all year round—why should they perform those (charitable) acts? Why should they perform charitable acts, and not those acts that provide them greater EU?

At this stage, I introduce a distinction between the *strong sense of mutual advantage* and the *weak sense of mutual advantage*. The weak sense of mutual advantage appeals to symbolic utility, while the strong sense of mutual advantage appeals to expected utility. Mb(CM)A endorses the latter, and claims that we are rational as long as we engage in participatory social activities that maximize expected utility, but Mb(DV)A endorses the former, and claims that we are rational to the extent we engage in participatory social activities that maximize symbolic (factoring EU as well). By appealing to EU Mb(CM)A rules out unilateral performance of actions, just in case they do not maximize EU. And by

ruling our unilateral performance of actions it fails to recognize the distinction between the weak and strong sense of mutual advantage.

Beyond EU-benefits, we can think of SU-benefits as well. Suppose I pride myself in being a good neighbor, then I would be disposed to do things that express this value I hold about myself. We might even suppose that I am an SU-seeker and I seek out actions that express this value. On this account, my doing things for my neighbors need not have additional EU benefit—I am not just an EU-seeker. Suppose you are my neighbor, and say you are sick, or just had a baby, then given my value or belief about good neighborliness, it is reasonable to suppose I will be moved to mow your lawn, even if that means I am going to miss out of the once-in-a-life time music concert. Your lawn might have grown beyond the limit allowed by the city. On this account, my mowing your lawn satisfies the weak sense of mutual advantage, even though it fails the strong sense of mutual advantage, for in your absence I would not have these benefits.[360] Mother Theresa or the moral philanthropist might still perform charitable acts, even if their EU-seeking interests are diminished or almost extinguished, just in case those acts express for them values that they share and care about; it is good to be compassionate, to be generous, to be a humanitarian.

[360] Could this lend itself to the broader argument that the rich is *obliged* to provide for the poor because they derive some utilities from the very relationship? The absence of the poor would extinguish these benefits. Do the famous not need the infamous to perpetuate their fame? If so, could they be *obliged* to 'pay' for 'using' the poor to perpetuate their fame? And if they refused to pay are they not using the poor as means (or mere means) to their end or at the very least 'parasiting' on the poor. Robert Frank who Braybrooke refers to in "Social Contract Theory's Fanciest Flight" has elegantly and profoundly made a similar point. Summarizing the point about parasitism that Frank makes in his book, *Choosing the Right Pond* (New York: Oxford University Press, 1985), which was published only after Gauthier had finished MbA Braybrooke says, "…superior status is a good which people are prepared to pay for—and prepared to relinquish only when given compensation. If those who succeed in the market get superior status with their wealth yet fail to compensate those who with less success are to be content with inferior status, they will be parasites themselves," p. 759.

Were we to limit the benefits people get from unilateral performance of actions to s symbolic utility, there is no reason to think that SU benefits are insufficient to motivate them to choose cooperation either in situations of diminished or almost extinguished expected utility. I can still mow your lawn in the face of diminished or extinguished expected utility insofar as mowing your lawn expresses for me the value of good neighborliness. An act, for an SU-seeker can still be performed even when the action's instrumental value is believed nil, or in case of the desire for revenge, where achieving this desire symbolically is better than inflicting an actual danger.

By recognizing the value acts have for people, Mb(DV)A recognizes the full range of reasons for acting. Given the expected utility and symbolic utility dimension of acts, it is easy to understand how by favoring non-secession acts over secession acts, better-off members benefit themselves at the very same moment they benefit less well-off members. If acts stand for something for a decision-maker, in addition to the consequences of those acts, then this expressiveness would be included in that decision-maker's reasons for acting. If we suppose that choosing non-secession acts stands for 'good cooperators' for productive members, the same way that not burning the flag represents good citizenship for a person who chooses the 'no-flag burning act,' then it is the case that productive or better-off members perform a unilateral act when they decide to support the North in their unproductiveness. But in performing a unilateral act better-off members benefit themselves at the very same moment they benefit unproductive and less well-off members. On this account, morality can be said to be mutual, and cooperation

understood thus is symbolically and intrinsically positive, even when it may or may not be instrumentally positive.

Bibliography

Baier, Annette, "Pilgrim's Progress," in *Canadian Journal of Philosophy*, vol. 18, June 1988, pp. 315-330.

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

Braybrooke, David, "Gauthier's Foundations for Ethics under the test of Application," in *Contractarianism and Rational Choice*: *Essays on David Gauthier's* ***Morals by Agreement***, Peter Vallentyne (ed.), New York, Cambridge University Press, 1991.

_____, "Social Contract Theory's Fanciest Flight," *Ethics*: *An International Journal of Social, Political, and Legal Philosophy*, vol. 97, no 4, July 1987.

Castiglione, Dario, "History, Reason and Experience: Hume's Arguments against Contract Theories," in *The Social Contract from Hobbes to Rawls*, David Boucher and Paul Kelly (eds.), London, Routledge.

Cooper, Wesley "Nozick, Ramsey, and Symbolic Utility," *Utilitias*, Vol. 20, no. 3, September 2008.

Copp, David, "Contractarianism and Moral Skepticism," in *Contractarianism and Rational Choice*, Peter Vallentyne (ed.).

Danielson, Peter, "The Visible Hand of Morality: Review of David Gauthier's *Morals by Agreement*," in *Canadian Journal of Philosophy*, vol.18, no.2, June 1988, pp.357-384.

DeBruin, A. Debra, "Can One Justify Morality to Fooles?" in *Canadian Journal of Philosophy*, vol. 25, no 1, March 1995, pp1-31.

Frank, Robert, *Choosing the Right Pond*, New York: Oxford University Press, 1985.

Freeman, Samuel "Introduction: John Rawls—An Overview," in *The Cambridge Companion to Rawls*, Samuel Freedman, (ed.), Cambridge, Cambridge University Press, 2003.

Harsanyi, John, "Morality and the Theory of Rational Behavior," in *Utilitarianism*

and Beyond, Amartya Sen and Bernard Williams (eds.), Cambridge, Cambridge University Press, 1982.

Gauthier, David, "The Social Contract: Individual Decision or Collective Bargain?" in *Foundations and Applications of Decision Theory*, Vol. 2, C.A. Hooker, Jim Leach, and Edward McClennen (eds.) Dordrecht, Holland, D. Reidel,1978.

_____, "Moral Artifice: A Reply by Gauthier," in *Canadian Journal of Philosophy*, vol. 18, 1988, pp. 385-418.

_____, *Morals by Agreement*, Oxford, Oxford University Press, 1986.

_____, "Morality, Rational Choice, and Semantic Representation: A Reply to my Critics," in *Social Philosophy and Policy*, vol. 5, 1988, pp. 172-221.

_____, "Justice as Social Choices," in *Social Contract Theory*, Michael Lessnoff (ed.), Oxford, Basil Blackwell, 1990.

_____, *Moral Dealing: Contract, Ethics and Reason*, Ithaca and London, Cornell University Press, 1990.

_____, "Why Contractarianism?" in *Contractarianism and Rational Choice*, Peter Vallentyne (ed.), 1991.

_____, (1991), "Rational Constraint: Some Last Words," *Contractarianism and Rational Choice*, Peter Vallentyne (ed.), 1991.

_____, Uniting Separate Persons," in *Rationality, Justice and the Social Contract: Themes from* ***Morals by Agreement***, Gauthier and Robert Sugden (eds.), Ann Arbor, University of Michigan Press, 1993.

_____, "Hobbes Social Contract," in *The Social Contract Theorists: Critical Essays on Hobbes, Locke, and Rousseau*, Christopher W. Morris (ed.), Rowman & Littlefield Publishers, Inc., New York, 1999.

Hampton, Jean, "Can We Agree on Morals?" in *Canadian Journal of Philosophy*, vol.18, June 1988, pp.331-355.

_____, "Comments on Gauthier's Hobbes's Social Contract," *Noûs*, vol. 22, March 1988, pp.85-86.

_____, (1991), "Two Faces of Contractarian Thought," in

Contractarianism and Rational Peter Vallentyne (ed.), 1991.

_____, "Equalizing Concessions in the Pursuit of Justice: A Discussion of Gauthier's Bargaining Solution," in *Contractarianism and Rational Choice*, Peter Vallentyne (ed.), 1991.

Harman, Gilbert, *The Nature of Morality*, New York, Oxford University Press, 1977.

_____, "Rationality and Agreement: A Commentary on Gauthier's *Morals by Agreement*," in *Social Philosophy and Policy*, vol. 5, 1988, pp. 1-16.

Heath, Joseph, "A Multi-Stage Game Model of Morals by Agreement," *Dialogue: Canadian Philosophical Review*, vol. 35, no. 3, Summer 1996, pp. 529-552.

Hobbes, Thomas, *Leviathan* (1651), Richard Tuck (ed.), Cambridge, Cambridge University Press, 1996.

Hubin, Donald and Mark B. Lambeth, *"Providing for Rights*," in *Contractarianism and Rational Choice,* Peter Vallentyne (ed.), 1991.

Hume, David, *A Treatise of Human Nature* (1739-40), L.A. Selby-Bigge (ed.) and notes by P.H. Nidditch, Oxford, Clarendon Press, 1978.

_____, *Enquiry Concerning the Principles of Morals,* 1751.

James, William, "The Will to Believe," in *The New World*, June 1896.

Kant, Immanuel, *Groundwork of the Metaphysics of Morals*, 1785.

_____, *Metaphysics of Morals* (1797), in ***Practical Philosophy***, Mary J. Gregor, (trans.), Cambridge, Cambridge University Press, 1996.

_____, "We Have No Duties to Animals," from *Lectures on Ethics*, trans. Louis Infield, London, Methuen Press, 1930.

_____, *Perpetual Peace: A Philosophical Sketch* (1795), in ***Kant's Political Writings***, H.S. Reiss (ed.) Cambridge, Cambridge University Press, 1991.

_____, "On the Common Saying: That May be Correct in Theory, but it is of No Use in Practice."

Kavka, Gregory, *Moral and Political Theory,* Princeton, Princeton University Press, 1986.

Kojeve, Alexandr, Introduction *to the Reading of Hegel*, Allan Bloom (ed.) and trans. by J. H. Nichols, Jr., New York, Basic Books, 1969.

Kymlicka, Will, "Justice and Minority Rights," in Robert E. Goodin and Philip Pettit (eds.) *Contemporary Political Philosophy: An Anthology*, Oxford, Blackwell, 1991.

_____, Contemporary *Political Philosophy: An Introduction*, second edition, Oxford, Oxford University Press, 2002.

Levine, Andrew, *The General Will: Rousseau, Marx, Communism*, Cambridge, Cambridge University Press, 1993.

Locke, John, *Two Treatise of Government* (1689), Peter Lasltett (ed.), Cambridge, Cambridge University Press, 1988.

_____, *Second Treatise of Government*, 1690.

Macintosh, Duncan, "Two Gauthiers"? *Dialogue: Canadian Philosophical Review*, vol. 28, 1989, pp.43-61.

MacIntyre, Alasdair, *After Virtue*, Notre Dame, IN, University of Notre Dame Press, 1981.

_____, *Against the Self-Images of the Age*, Notre Dame, University of Notre Dame Press, 1978.

_____, *Whose Justice? Whose Rationality?* Notre Dame, University of Notre Dame Press, 1988.

Mackie, John L., *Ethics: Inventing Right and Wrong*, Harmondsworth, Penguin, 1977.

Nagel, Thomas, *The View From Nowhere*, New York, Oxford University Press, 1989.

_____, "Rawls and Liberalism," in *The Cambridge Companion to Rawls*, Samuel Freedman, (ed.), 2003.

Narveson, Jan, "Gauthier on Distributive Justice and the Natural Baseline," in *Contractarianism and Rational Choice*, Peter Vallentyne (ed.), 1991.

_____, *The Libertarian Idea,* Ontario, Broadview Press, 2001.

Nietzsche, Friedrich, *On the Genealogy of Morals*, 3rd edition, translated by Walter Kaufman and R.J. Hollingdale, New York, Random House, 1967.

Nozick, Robert, "Newcomb's Problem and Two Principles," in *Essays in Honor of C. G. Hempel*, N. Rescher et al (eds.), Dordrecht: Reidle, 1969.

_____, *Anarchy, State and Utopia,* New York, Basic Books Inc., 1974.

_____, *The Nature of Rationality*, Princeton, New Jersey, Princeton University Press, 1995.

Parfit, Derek, *Climbing the Mountain* (new book manuscript).

Paul, Ellen Frankel, Fred D. Miller & Jeffrey Paul (eds.), *The New Social Contract: Essays on Gauthier*, NY, Blackwell, 1988.

Rainbolt, W. George, "Gauthier on Cooperating in Prisoner's Dilemmas," *Analysis*, vol. 49, October 1989, pp. 216-220.

Rawls, John, *A Theory of Justice*, Cambridge, Harvard University Press, 1999.

_____, *Political Liberalism*, New York, Columbia University Press, 1993.

Ripstein, Arthur, "Gauthier's Liberal Individual," *Dialogue: Canadian Philosophical Review*, vol. 28, 1989, pp. 63-76.

Rousseau, Jean-Jacques, *Discourse on the Origin and Basis of Inequality Among Men*, 1754.

_____, *Of the Social Contract*, 1762.

Sayre-McCord, Geoffrey, "Deception and Reasons to be Moral," in *Contractarianism and Rational Choice*, Peter Vallentyne (ed.).

Schwartz, David T, "The Limits of Self Interest: An Hegelian Critique of Gauthier's Compliance Problem," *Southwest Philosophy Review: The Journal of the* Southwestern Philosophical Society, vol. 13, January 1997, no. 1, pp. 137-146.

Shafir, Eldar and Amos Tversky, "Thinking through Uncertainty: Nonconsequential Reasoning and Choice in *Preference, Belief, and Similarity*, Eldar Shafir (ed.), Cambridge, MIT Press, 2004.

Singer, Peter, "Famine, Affluence, and Morality," in *Ethics in Practice: An Anthology*, Hugh Lafollette (ed.), Oxford, Blackwell, 1997.

Smith, Holly, "Deriving Morality from Rationality," in *Contractarianism and Rational Choice*, Peter Vallentyne (ed.).

Sobel, J. Howard, "Interaction Problems for Utility Maximizers," *Canadian*

Journal of Philosophy, vol. 4, June 1975, pp. 677-688.

_____, "Constrained Maximization," in *Canadian Journal of Philosophy*, March 1991, pp.25-51.

_____, "Straight Versus Constrained Maximization," in *Canadian Journal of Philosophy*, vol. 23, no. 1, March 1993, pp. 25-54.

Sugden, Robert, "The Contractarian Enterprise," in *Rationality, Justice and the Social Contract: Themes from* ***Morals by Agreement***.

Suits, Bernard Herbert, *The Grasshopper: Games, Life and Utopia*, Toronto, Broadview Press, 2005. Original edition published in 1978.

Taylor, Charles, *Philosophy and the Human Sciences: Philosophical Papers 2*, Cambridge, Cambridge University Press, 1985.

Vallentyne, Peter, "Contractarianism and the Assumption of Mutual Unconcern," in *Contractarianism and Rational Choice*, Peter Vallentyne (ed.), 1991.

Voice, Paul, *Morality and Agreement: A Defense of Moral Contractarianism*, New York, Peter Lang, 2002.

Walzer, Michael, *Spheres of Justice*, Oxford, Blackwell, 1983.

Williams, Bernard, *Ethics and the Limits of Philosophy*, Cambridge, MA: Harvard University Press, 1985.

Wolff, Jonathan, An *Introduction to Political Philosophy* (revised edition) Oxford, Oxford University Press, 2006.

Yi, Byeonguk, "Rationality and the Prisoner's Dilemma in David Gauthier's ***Morals by Agreement***," *Journal of Philosophy*, vol. 89, no. 9, September 1992, pp. 484-495.