TRACTABLE CONTRACTARIANISM

by

Christopher Miles Tucker

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

Waterloo, Ontario, Canada, 2001

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

Abstract

Contractarians are interested in producing a fundamental justification for moral or political institutions. A fundamental justification must both produce: i) compelling reasons to adopt, and ii) rely on grounds other than, the moral theory to be so justified. Taking instrumental rationality to be independent of any moral theory and sufficiently likely to lead to action, contractarians argue that a moral or political system is justified by its instrumental value to agents.

Contractarians have previously relied on a bargaining model of contractarian justification, where agents to be bound by a particular set of moral dictates are understood as if they are bargaining with each other over which rules to adopt. I find this unacceptable: no agent could foretell any other agent's worth to him or her in the future – which is critical in bargaining theory. Agents do not have prior knowledge regarding which rules will thereafter govern their interactions, and therefore have no way to reliably estimate the ability of each to enhance or detract from each other's preferences.

I suggest an alternative. Summing each person's preferences over each set of rules which might govern interaction, we identify the set of rules which is thereby ranked highest to be justified. This process is also found to be calculable in principle (tractable), which is a necessary condition for providing a fundamental justification. Nevertheless we do not have sufficient information at this time to proceed. We do not have access to each person's preferences and beliefs regarding the set of moral systems. Exploring the literature surrounding contractarianism leads us to recommend an examination of psychology and economics to attempt to identify for each agent preferences that are of paramount import to that agent.

Acknowledgements

I would first like to thank my supervisor, Jan Narveson. His tireless efforts, boundless enthusiasm, and critical insights have inspired me and rendered me entirely in his debt. William Abbott and Dave DeVidi have also been amazingly supportive and giving of their time and opinions – even when time was in particularly short supply. A special thanks goes out to Paul Thagard, without whom I would have never investigated the computational complexity of contractarianism.

More generally I would like to thank the department of philosophy at the University of Waterloo. The congenial atmosphere and lively philosophical discussions - to be had for the price of merely knocking on a door - kept philosophy out in the open, instead of hidden behind computer screens. I would also like to thank Debbie Dietrich and Linda Daniel, both of whom kept me informed about what I was supposed to be doing, and the best way to go about doing it. I am also grateful to both of them for countless things that they did behind the scenes (which still remain a mystery to me) to ensure that my years here passed as smoothly as they did.

To my fellow graduate students, I thank them for listening to me when I prattled on about contractarianism, and acted as if I was saying something interesting, or at least intelligible. Most especially I thank Christine Freeman, whose friendship I cherish. She kept me aware of the social world around me, and even occasionally of my professional obligations. Thanks for saving me from myself.

And lastly, to Susan Dimock. Her love and support sustains me, and her encouragement inspires me to believe.

Table of Contents

Introduction

I. The State of Moral Debate

Patrick and James are having a dispute over the propriety of taxing the better off for the purposes of redistributing the revenues gathered to the less fortunate. Patrick says, "Look, James, it is clear that we all have a duty to safeguard the dignity of each person. Being extremely poor leads to all kinds of indignities – people don't want to even look at you if you're poor, children mock you and you have to resort to extremely degrading behavior in order to survive. Since we have a duty to protect the dignity of people, and since being extremely poor is degrading, we must gather resources from each better off person, and alleviate the indignities suffered by those worse off."

James shouts, "I'm not even going to bother cataloguing the number of problems with that supposed argument! The whole argument is absurd on its face, because we do not have a duty to safeguard people's dignity! All we have to do is let people do what they want with the stuff that they own as long as it doesn't impede the abilities of others to do the same!" Pausing to regain his composure, he then continues, "Patrick, please get serious. You can't justify theft by saying you don't want to see poor people treated in certain ways – we don't have a duty to maintain people's dignity, you can't prove to me that we do, and the reliance upon this duty is too convenient. It proves only what you want it to prove, and you can provide me with no reasonable proof that such a duty exists."

Patrick replies, "Well, maybe I can't, but you can't prove to me that there isn't such a duty."

James shoots back, "I don't need to prove it to you. The burden of proof is surely on the person who wants me to constrain my behaviors in certain ways."

Patrick yells, "Well then you can't prove that I ought to refrain from taking what you own, and giving it to someone else!"

James yells, "Well it's just obvious, you idiot!"

* * * * * * *

Kathryn and Anne are having a disagreement regarding whether or not Anne did Kathryn wrong. Kathryn wails, "Anne, how could you! You promised to help me study for my exam, and then you didn't show up! There's only an hour left and I'm going to fail for sure now."

Placatingly, Anne responds, "Look, Kathryn, I'm sorry that you're going to fail your exam, but I didn't have a choice. On my way to your house, I saw an elderly man looking disoriented. He almost wandered into traffic! I had to assist him – I walked him to his house, and got him inside, out of the sun. Who knows what might have happened had I not given him a hand?"

Kathryn, not placated in the least, replies, "It doesn't matter what might have happened. You made a promise, and had to fulfill that commitment. You owed me!"

Anne replies, "Of course I owed you, but that changed as soon as I saw the man in distress. I had an obligation to you that was overridden by the obligation I have to give aid to those in immediate danger."

Kathryn shoots back, "I don't see how that's so clear!"

Anne replies, "Well now you're just being contrary! Look, I'm sorry that you're going to fail your exam, but you know that it's obvious that, morally speaking, I had no choice!"

* * * * * * *

Jan and Susan are having an argument about women's rights. Jan states, "Look, we all know that all people should be treated equally, and clearly women need to be treated as equals to men. We also know that in the past this hasn't happened. In order to correct this injustice, we have to engage in some form of affirmative action policy! I think that the most effective form of affirmative action is one that forces companies to fill a quota system!"

Susan responds, "Jan, have you taken leave of your senses? Of course I agree with everything that you have said, right up to the point where you suggest a quota system. However, the most effective form of affirmative action is clearly one that requires only that the number of applicants interviewed be composed equally of women and men."

Jan responds, "Look, the studies that I have right here clearly show that there is no basis for your claim..."

Interrupting, Susan exclaims, "Jan, those studies are a load of crap! Not one of them makes use of a reliable method. Clearly the studies that have been reputably carried out show ..."

(And the struggle for women's rights continues...)

II. The Problem

Moral debate does not usually result in either a clear victor or consensus. Reliance upon basic moral intuitions to derive a favored conclusion has become the mainstay of ethical debate. Unfortunately, we do not all share the same basic moral intuitions, and the moral intuitions that we do share do not usually all recommend a unique moral course of action. If either of these possibilities were the case, we would not recognize that a problem exists. But they are not, and so we recognize a problem in ethical theory – mere reliance upon basic moral intuition does not provide reliable recommendations regarding what it is that we ought to do.

Overcoming this problematic result of the current status of the debate requires first the justification of a particular set of moral rules (a moral theory). Given such a justification we could conclude that the theory is more than simply a set of moral intuitions. For this justification to have the effect on moral debates that we wish, we would all have to agree that this justified set is the one that should be decisive in recommendations regarding moral action. For everyone to come to agree regarding the practical applicability of a particular set of moral rules, each person would have to be persuaded to find this set superior both to their intuitions and all other competitor sets of rules. We would have to provide compelling reason for its adoption.

By producing such a justification we would, in effect, be providing a set of rules against which we could check moral intuitions to see if they have been borne out by the justification. If a moral intuition regarding the duty to refrain from X was on the list of moral behaviors (one of the things agents have to do is refrain from X), then that intuition would be justified. To come to agree regarding any course of action, we would also have

to come to agree about what is morally required of us, which would sometimes require that we agree about certain empirical matters. This latter issue is not what we here choose to focus upon. We instead wish to describe and defend a particular method that would most reliably result in a justification of a moral theory: instrumental contractarianism.

An instrumental contractarian account attempts to justify a moral or political theory by way of reference to individual people's preferences and beliefs. It attempts to provide good reason for accepting the dictates of moral theory, which are social constraints on individual behavior, by relying on individual beliefs and preferences.[1] Given each person's preferences and beliefs, the contractarian concludes that each person's interests are best served, all things considered, by a particular moral theory. The contractarian holds this theory to be justified.

Wishing to suggest that a particular version of this project is the most plausible method by which to provide a fundamental justification of a moral theory, we first critically survey the field of popular alternatives to show both what contractarianism is, and what it can accomplish that these contenders cannot. We then argue that a particular form of instrumental contractarianism (amalgamative contractarianism) is the most plausible method by which to provide a justification by showing that the popular alternative contractarian account (market contractarianism) is implausible as a method of producing a social choice via individual choices. According to amalgamative contractarianism, the calculation which identifies the set of rules that is justified is

[1] Although a perspicuous way to describe our efforts here, it will become clear over the course of this project that we think it is misleading to describe moral rules as 'constraints'.

sufficiently simple to calculate that we could suppose that such a production is possible in a timely fashion. Without timely output, no justification could further moral debate.

However simple the contractarian calculation is found to be, we also find that the necessary inputs are lacking. In order to produce a maximally compelling justification, we must make use of the complete sets of each person's preferences and beliefs – in no other way can we persuasively argue that each person has the best possible reason to accept the justification presented. We do not have access to such information, however, and it is not foreseeable that we will have such access in the near future. Finding the furtherance of moral debate sufficiently important to not be satisfied with letting matters remain as they currently are until such time as such complete sets of preferences and beliefs may be provided for our use, we ask what alternatives we are left with.

Finding one promising avenue of exploration in the possible production of a partial list of preferences and beliefs, we examine what properties such a list must have in order to have a chance at providing each agent with, if not the best possible reasons, at least the best reasons possible *right now* for the adoption of a particular moral code. It was a skepticism regarding the possibility of being able to provide such a justification that led to the mere intuition clashing that typifies moral debate today. It is a skepticism regarding the possibility that such intuition clashing will ever allow moral debate to progress that forces us to re-examine the possibility of providing a timely justification. It is on this possibility that we now focus.

CHAPTER 1
~ These Things I Don't Believe ~

This chapter will be dedicated to both defining and justifying instrumental contractarianism through contrast with the alternatives to it. I will first distinguish instrumental contractarianism from other forms of contractarian, or contractualist, justifications. This will primarily involve an exploration of John Rawls' *A Theory of Justice*.[2] Secondly I will distance instrumental contractarians from those theorists who propose an 'idealized dialogue' approach to moral/ political justification. John Rawls' *Political Liberalism*[3] and Jürgen Habermas' *Moral Consciousness and Communicative Action*[4] will serve as examples of the idealized dialogue approach. Finally, Peter Danielson's *Artificial Morality*[5] will serve as a focus through which to distinguish instrumental contractarianism from evolutionary theories of rational behavior.

This chapter should not be understood purely, or even primarily, as an aid in identifying an instrumental contractarian account. The general approaches examined here will be found unable to produce a compelling justification of a moral or political system. As each of the particular presentations examined will be found lacking in some fairly basic way, we seek to explain not only what we are not endorsing, but also why we are not endorsing it. Taking the inadequacies of these various theories to heart will then make our aim of presenting an instrumental contractarian account that is capable of producing a fundamental justification much easier. It is to the extraction of such lessons that we proceed.

---

[2] John Rawls, *A Theory of Justice*, Harvard University Press, (Cambridge, Massachusetts), 1971.
[3] John Rawls, *Political Liberalism*, Columbia University Press, (New York, New York), 1996.
[4] Jurgen Habermas, *Moral Consciousness and Communicative Action*, MIT Press, (Cambridge, Massachusetts), 1990.
[5] Peter Danielson, *Artificial Morality*, Routeledge (New York, New York), 1992.

I Contractualism

Both contractarian and contractualist attempts to justify any moral or political system are grounded in a particular theory of rational agency. The specific understanding of "rational agency" and "rational agents" that both contractarians and contractualists begin with is, roughly speaking, one which accords with the economic understanding of those terms. An economically rational agent acts in such a way as to attempt to maximize the satisfaction of her subjective preferences, given her information about both those preferences and the situations in which she finds herself. Given general information about the world, and in particular facts about strategic interaction, these theorists attempt to provide an account of why these agents would find it rational to endorse certain rules constraining their behavior in various ways. Relying on the rationality of the dictates presented, such theorists attempt to provide reasons for their adoption, reasons that, they suppose, would prove compelling. This compulsion is, as outlined above, a necessary condition of any successful justification. Contractualists, by way of contrast with contractarians, modify the psychology of the agents being examined. Typically they do this by either including or excluding certain types of preferences for reasons other than that these preferences could or could not be expected to be found in all rational agents. They do so at the peril of their enterprise, as we shall soon see.

I.i A Theory of Justice

John Rawls' *A Theory of Justice* has, as its main goal, the justification of our most basic intuition about justice: that justice overrides claims about social welfare. The basic rights of the individual cannot be violated for any increase in social benefit. Leaving

aside Rawls' definition of utilitarianism, which serves as the theoretical champion for advancing social welfare over the concerns of the individual, we will focus upon the two other necessary components of this argument. For his argument to be compelling, Rawls both must present a convincing definition of justice, and produce good reasons why society should choose to champion justice over social welfare. We will now focus upon the argument Rawls provides to suggest why society should choose to champion justice over social welfare.

In order to justify society's championing justice over the advancement of social welfare, Rawls turns to contract theory. In presenting a contractarian account, the theorist enjoins us to imagine an hypothetical situation in which people come together prior to their being constrained by any social regulation. These persons are to then decide which principles would best regulate the various competing claims to social benefits that particular persons are likely to demand as their privilege. These rational agents who have managed to come to agreement over which principles will thereafter arbitrate between competing claims over social benefits are supposed to have identified principles that would regulate a society as if it were a voluntary association. As these rational agents are to be understood as rational representatives of actual people, the resulting 'contract' comes as close as possible to expressing what we would have agreed to be bound by were we to have had a choice in the matter. In a sense, therefore, one might say that society endorses these principles of regulation. One could further explain why society endorses these principles; society endorses these principles because it has found them to be in their interests.

In order to identify as precisely as possible those principles which people would consent to be bound by thereafter, it is clear that these rational agents ought to resemble as closely as possible the actual people to which these principles are to be said to apply. The result of this construction, however, would likely include *all* of the rules that people would agree to be bound by. This list could include rules of etiquette, for instance. This is not acceptable, given Rawls' project. Rawls is not interested in identifying all of the principles governing social interaction; he is interested is justifying the precedence of justice over utilitarian concerns. In order to do this, he must identify the principles of justice, and those principles alone. Harkening back to our considered convictions about justice, Rawls suggests that we can all agree that if any set of principles are to be identified with justice, they must be ones that do not produce any decisions over competing claims based on differences between the claimants that are arbitrary from a moral point of view.[6]

Suggesting what he takes to be quite basic examples of what we would all agree are arbitrary social contingencies about people – their natural talents, their social status from birth, etc. – Rawls systematically denies his rational agents epistemological access to these contingencies. Supposing that rational agents with this information would bargain in ways thought beneficial to their contingent state of affairs, and that this would thereby introduce arbitrary distinctions into the deliberations about which principles that society would endorse, he denies that a process including such information could identify principles of justice. The principles of justice must be identified via a procedure that makes no appeal to these social contingencies, a procedure thereby more likely to identify

[6] Rawls identifies considered convictions as those evident in our "judgments in which our moral capacities are most likely to be displayed without distortion." [p. 47]

principles that are obviously principles of justice. Agents are to bargain as if they had been placed under a 'veil of ignorance' with regards to these contingencies. These principles of justice are then compared with the principles of utilitarianism, and the case for the precedence of the former presented.

It is obvious that this procedure does take us some way from identifying the principles that actual people would have endorsed had they the chance, but it does so in a way that Rawls finds unobjectionable. Given that these modifications have been introduced due to our most settled convictions about justice, convictions that we all share and make without hesitation, Rawls supposes that the results of this theory are still ones that we can be said to accept. The conclusions of this theory are to be understood as resulting from our interests combined with our most significant sentiments about justice.

The original position, and the veil of ignorance that primarily characterizes it, are both the results of what Rawls takes to be our most basic intuitions about what is reasonable to require in order to assure that the principles proposed a) are principles of justice, and b) are rationally acceptable. We turn first to the veil of ignorance. Under a veil of ignorance agents are supposed to have no access to arbitrary information, information that would taint the resulting inquiry into the correct principles of justice. This entails

> [T]hat ... no one knows his place in society...nor does he know his fortune in the distribution of natural assets and abilities...Nor...does anyone know his conception of the good...or even the special features of his psychology such as his aversion to risk ... The parties do not know the particular circumstances of their own society. ... The persons in the original position have no information as to which generation they belong.[7]

---

[7] *Theory of Justice*, p. 137.

This, naturally enough, raises the question of what information the parties in the original position are supposed to have. Together with an index of goods that it is supposed every rational agent would want more rather than less of, given that they have any rational plan of life at all, Rawls suggests that:

> [T]he only particular facts which the parties know is that their society is subject to the circumstances of justice and whatever this implies. It is taken for granted...that they know the general facts about human society. They understand political affairs and the principles of economic theory; they know the basis of social organization and the laws of human psychology. Indeed, the parties are presumed to know whatever general facts affect the choice of the principles of justice. There are no limitations on general information.[8]

This ensures that, in spite of the previous extreme curtailing of information that agents may access, they still have both sufficient information and motivation to come to agreement.

The parties in the original position are further assumed to have an equal say in the proceedings, to not suffer from envy, more generally not to take an interest in one another's interests (mutually unconcerned), and to be capable of effectively constraining their actions to accord with the dictates of a reasonable conception of justice and be known (by others) to be so capable. Their equality is demanded by our intuitions about justice, as an unequal say is inherently unjust and any procedure that allows inequality to affect the output will not identify justice. Envy tends to make agreement less likely, and is therefore excluded. Positive fellow feelings would tend to render the project irrelevant; beings concerned with each other may be supposed to be less competitive.[9] Finally, the

---

[8] *Theory of Justice*, p. 138.

[9] Of course, there is another, less charitable explanation for Rawls' exclusion of positive fellow feelings from the original position. The inclusion of such feelings would make the Utilitarian's case that much easier to make, since each person would have a motive to act from other peoples' interests.

capacity to adopt a normally effective sense of justice is assumed in order to make the procedure meaningful; there is no point in attempting to agree to a set of principles that no one would thereafter be constrained by. It is under these restrictions that Rawls concludes agents must conduct their comparison of a society regulated by utilitarian concerns with a society regulated by constraints of justice.

The particular constructions of the principles of justice that Rawls identifies, and the arguments forwarded to justify the priority of his principles of justice, are not relevant for our purposes. His method of constructing the choice situation is faulty, and we need discuss no more in order to take issue with it. His argument is fatally biased. We shall argue that it is clearly impossible to suppose that constructing the choice situation so that it accords with our sense of justice would not irreparably bias the argument concerning the priority of justice over utilitarian concerns. It is the further elucidation of this concern to which we know turn.

I.ii The Role of Identification in Justification

Rawls suggests that many of the conditions constraining the original choice position and the veil of ignorance are 'reasonable' to impose on a theory of justice, and therefore on any theory arguing *for* a conception of justice. No theory of social interaction that allowed for decisions among competing claims to be made on attributes that are arbitrary from the moral point of view could be called a theory of justice; from this Rawls justifies imposing the veil of ignorance. No theory that allows for the unequal treatment of a society's members could be called a theory of justice; and so each person is supposed to have an equal say. No theory that fails to account for the interests of future generations could be called a theory of justice; and so each person is assumed to

have an interest in the interests of future generations. Taking these and other sentiments to be our settled convictions about what a theory of justice must recommend, whatever else it recommends, Rawls concludes that each of these sentiments justify a modification of the method whereby he seeks to identify and justify his favored principles of justice. We will leave aside a quite serious reservation about whether, in fact, 'we all' share these convictions. What is of concern is the reasoning behind justifying a theoretical modification to a justificatory enterprise because of constraints upon the *output* of that theoretical justification.

The modifications in this particular project render the outcome of Rawls' arguments unsurprising and not compelling. Arranging the choice situation so that it closely mimics our most basic intuitions about justice and the agents in such a way that they are fairly assured of choosing just institutions, and then showing that it is reasonable to suppose that these agents choose principles of justice over utilitarianism, is not showing very much at all.[10] The case is clearly arranged in such a way as to load the dice from the start. One must always take care that if the dice are loaded, that they are loaded to the benefit of the opposition. It would have been far more satisfactory to construct an argument that assumed that agents had strong positive fellow-feelings, and were average utility maximizers in all choice situations, and that these agents *still* chose to accept principles of justice over a utilitarian system of governance. As it is, the bias of the argument renders the conclusions questionable.

---

[10] G. E. Pence takes notice of many of Rawls' question-begging devices, both in his veil of ignorance and his formal constraints upon acceptable principles, in "Fair Contracts and Beautiful Intuitions". This can be found in pp. 137-152 of *New Essays on Contract Theory* (*Canadian Journal of Philosophy*, Supplementary Volume III, Kai Neilson and Roger A. Shiner, eds., 1977.)

More generally, however, one must be cautious about attempting to present a justification for a position which is modified in order to ensure that what is justified is that position. If one wants a justification to do any work at all, it must be constructed on grounds other than those one is attempting to justify. A justification of justice that entirely relies on that conception of justice does no work. "A therefore A" is uninteresting. We are not suggesting that Rawls' account is completely question begging. But, similarly to the point above, the degree to which one relies upon the concept to be identified has a direct relation to the degree to which the justification is uninteresting – the degree to which it is not a justification at all. Following Robert Nozick, a fundamental justification must appeal to none of the concepts of that realm.[11] It is this most desirable justification which must, ultimately, be attempted if the results are to be convincing. A justification of the particular theoretical construction making the case for principles of justice ought to make no reference to our intuitions about justice.

This is importantly different from suggesting that the argument constructed cannot lead one to inevitably choose one possible conclusion as justified over a field of contenders. We ought not, however, attempt to construct an argument in such a way so as to lead to a particular *preconceived* conclusion. A good argument ought not to leave one neutral regarding which possible conclusion to endorse, but it ought to convince us to adopt a particular conclusion from a basis of neutral (non-leading) premises. As simple as this point seems when stated clearly, parallel problems plague liberal studies because of a confusion between the two different types of neutrality. Liberals cannot claim that a just political system must have consequences that do not partially affect differing

[11] See *Anarchy, State, and Utopia*, Basic Books; New York, 1974, pp. 4-9.

conceptions of the good life, as some decisions will always have such an effect.[12] Liberal governments are, however, commonly criticized by other liberals because some policy or other has the effect of disadvantaging some particular conception of the good. Liberals can much more plausibly claim that, nevertheless, these policies must not have as their *aim* the advantaging or disadvantaging of some particular conception of the good.[13]

This is not to say that our most fundamental intuitions about justice have no place at all. Once a theory is constructed, and its justificatory force identified, and the argument concluded, the principles thereby justified ought to be contrasted with our most basic intuitions. If the output of the theory does not recommend equal treatment, then it is not a justification of a theory of justice. A theory that justified the better treatment of some based on their height cannot be said to justify a theory of justice. But this is entirely distinct from the construction of the justification. Given that our basic intuitions are to be made use of in identifying what it is that we have attempted to provide a justification of does not imply that we may make use of these notions when constructing the procedure by which the justification is attempted. Since this is the case, a fundamental justification of moral and political systems cannot make reference to moral and political values. Contractarians cannot, for example, assume equality of the worth of the participants if the assumption is based on moral considerations. Any assumptions must be drawn from a realm distinct from the moral and political.

---

[12] For example, choosing whether or not abortion is to be a legal medical procedure, or publicly funded, will inevitably result in some conception of the good (pro-choice, for example) being affected positively while another (pro-life) is affected negatively.

[13] Susan Dimock presents a discussion of these and related concerns regarding neutrality and liberalism in her "Liberal Neutrality", *Journal of Value Inquiry*, 34:2-3, 2000, pp. 189-206.

II Discourse Ethics

Turning now from non-fundamental attempts to justify moral and political systems, we proceed to investigate what I shall call, following Habermas, discourse ethics. These theories are attempts to justify strategic norms of interaction by appeal to the force of reasoned agreement; discussion among agents will lead to agreement as to the content of these norms. There is one feature in particular of which it is worthwhile to take note: agreement has binding force. Reasoned agreement produces a good reason to act on the dictates of these agreements. Accepted reasons tend to constrain behavior, for accepting a new claim involves creating a new belief, and belief tends to modify action. In such a way, idealized discourse theory attempts to secure the motivation necessary for the resulting theory to be a justification instead of a mere explanation.

Jürgen Habermas and John Rawls are the two theorists who I take to be the most famous proponents of such a view. Habermas' *Moral Consciousness and Communicative Action* and Rawls' *Political Liberalism* are both clear examples of this method. They also, between them, manage to chart the whole range of possible discourse ethics. Habermas suggests a theory that proposes that minimally idealized agents (that is, agents who are not idealized at all) are the correct ones with which to work; anything short of this necessarily diminishes the internal force of the argument. Insofar as an agent is conceived as separate from an actual discourse, it is the voice of the theorist that enters the picture, a voice which provides no further benefit to the discussion, and lessens the merit of the enterprise. So much the worse if the agent is not even "discussing" problematic norms. This is a critique to which we are sympathetic. So, we join Habermas in critiquing Rawls' *Political Liberalism*, in which the discourse situation

appears to be merely a vehicle for various theorists' discussion over the proper content of the political norms in question. In this section we will ignore the discussion of *A Theory of Justice* above, and focus only upon those elements of *Political Liberalism* that pertain to Rawls' acceptance of a discourse ethic.

Habermas' attempt must not be supposed to escape unscathed. While we are sympathetic to his approach, we suppose, perhaps pessimistically, that the internal force of mere reason is insufficient to produce the practical force we find necessary for a justification of a normative system. Belief in certain principles does not sufficiently compel action; hypocrisy flourishes, and any who suppose otherwise are simply not in touch with the real world. While Habermas' project may end up being a necessary part of any justificatory enterprise of moral theory, it is not the largest or most significant part. We will first outline Rawls' proposal, and then will discuss Habermas' offering. With both of these theories in hand, we will then proceed to critique the entire methodology as a system of justification.

II.i Political Liberalism

Rawls takes great pains to point out that his aims in *Political Liberalism* are: to suggest that the concept of a justified constitutional democracy whose citizens embrace a plurality of world-views is coherent; to provide a liberal theory of justice from which to develop the most appropriate institutions for the achievement of the ends of a liberal democracy; and to show that such a political system could be implemented in our world and be stable over time. If these were his only aims, *Political Liberalism* would hold only tangential interest for our purposes here. But *Political Liberalism* also has normative ambitions. It derives a conception of justice that a plurality of citizens of a

democratic regime may thoughtfully and freely endorse. Rawls later suggests that the point of his enterprise is to achieve a consensus about the appropriate forms and of functions of political institutions in a constitutional democracy dedicated to upholding the equality of its citizens as persons.[14] The latter normative aim obviously rests in large part on the former; it is only by allowing that those to be subject to the coercive powers of the state could endorse such treatment that a liberal could justify any form of political institution. Insofar as one is already persuaded, then, of the validity of a liberal constitutional democracy, and the moral intuitions which are supposed by Rawls to underlie such a regime, one cannot help but be interested in the claim that *Political Liberalism* is a tool through which we theorists may identify (morally) appropriate political institutions.

In *Political Liberalism* the voluntary assent of each person to the proposed system of regulations is tightly linked to the question of stability. Consensus regarding the proper institutional form of a pluralist democratic regime is not only contingent in part upon the possible endorsement of each of its citizens, but also upon the structure and content of Rawls' derivation of the principles of justice. Finally, the most obvious way to attempt to prove that a notion is coherent is to spell it out without contradiction. Accordingly, Rawls' project is divided into two distinct parts: the spelling out of a liberal-democratic conception of justice and the principles by which its most appropriate institutional forms are to be identified, and his argument for expecting that a society that is well-ordered by such a conception will be stable. Given that we are here primarily interested in the form of Rawls' justificatory enterprise, we will not examine his discussion of stability over time. Whether or not his particular formulation is likely to be

[14] *Political Liberalism*, pp. 45-46, 300.

stable over time when he is concerned with his constructed citizens is not of primary import for our purposes. The project by which he proposes to derive appropriate principles of justice is our concern. Ultimately, we do not find this project acceptable. We render no opinion on the second project; for, finding the first lacking, we have no reason to continue to beat what will have been found to be a dead horse.

Rawls' insistence that his theory of political justice (justice as fairness) be acceptable to a multitude of people characterized by incompatible systems of belief and value (comprehensive doctrines) leads him to attempt an articulation of political justice that relies on no particular comprehensive doctrine. This in its turn, together with the obvious advantage obtained by not relying on any controversial premises - and the need to rely on *some* premises - leads him to construct his system as far as possible on the core commitments of a democratic society.[15] There are two such commitments of which he centrally makes use when discovering democratic justice: society is to be viewed as a "fair system of cooperation over time"[16] and people must be regarded as both free and of equal worth.

Rawls is quick to point out that "society as a fair system of cooperation over time"[17] is distinct from the coercive coordination of citizen's actions. Cooperation implies an acceptance of the publicly recognized rules of interaction. These restrictions must be acceptable to each reasonable person given that all would act accordingly. These

---

[15] Although Rawls himself is frustratingly vague about the characteristics of these "fundamental ideas seen as implicit in the public political culture of a democratic society" [p. 13], I take him to mean that these ideas could be commonly seen to motivate the various institutions and regulations without which no regime could reasonably be called democratic. Given the plethora of necessary and sufficient conditions proposed by various theorists to guide us in the proper use of this term, Rawls' casual derivation of these fundamental ideas seems simplistic at best.

[16] *Political Liberalism*, p. 14.

[17] *Political Liberalism*, p. 15. Hereafter this will be treated as a technical term, and the quotation marks will be dropped.

regulations must also be fair; they must assure that all who abide by these regulations appropriately benefit thereby. The purpose of these regulations is to achieve and preserve the just distribution of the benefits of society, as specified by a suitable conception of justice. Partly by reference to this definition of what a society is, Rawls defines the relationship between a person, a citizen and a party in the original position.

Before suggesting how people are to be viewed as free and equal, Rawls first presents us with his view of what underlies our "everyday conception of persons as basic units of thought, deliberation [and] responsibility."[18] This derivation from our (uncontroversial and unscientific) conception of a person in turn provides the building blocks from which he constructs his notion of the citizen as a political person. Ultimately, it is the citizen that he attempts to model in the original position, from which he derives his principles of justice (justice as fairness). But this is to anticipate.

Persons as political entities are identified as *at least* being entities that *may* be members of society; a person may be a fully cooperating member of society over a complete life. Given our common understanding of what it is to be a person, this suggests to Rawls that a person is to be viewed as both reasonable and rational, and as having an organized view of the world, with which she steers herself through it.[19] To be rational is to be, in a non-naive sense, an economic agent – to have goals and desires (a conception of the good) that one strategically acts upon given one's beliefs. To be reasonable implies two things: a desire to propose and discuss principles for cooperation justifiable to all, acting upon them when assured that others will too *and* that they

---

[18] *Political Liberalism*, p. 18n.

[19] More accurately, Rawls states that a person's being reasonable and rational is derived from our conception of persons as being responsible for their actions [p. 50]. A person's ordered world-view is an obvious requirement of his having thought/deliberative processes.

recognize that people inevitably disagree about their world-views, and so they cannot appeal to these views when justifying their proposed principles.

Citizens, by way of contrasting this concept with the concept of a person, are identified as entities that *are* fully cooperating members of society over a complete life. It is by possessing the characteristics of a person spelled out above to the "minimum degree necessary to be fully cooperating members of society" that people are conceived of as equal for political purposes; people are equal *as citizens*. This is, of course, only one half of the democratic conception of the political person. People are not only conceived of as equal, but also as free. Citizens are understood as free in three ways: they have a freestanding public identity, with rights and duties, that does not depend on their particular conception of the good; they are freestanding sources of legitimate claims against their government; and they are seen as freely endorsing their particular conception of the good. It is by reference to these attributes that Rawls concludes that we may conceive of citizens as also having a reasonable moral psychology.[20] To have a reasonable moral psychology is to have the ability to internalize a conception of justice and the motivation to act upon it; to do one's part too in a justly organized society given assurance that others will as well; and to develop stronger bonds of trust in one's institutions and fellows over time, given continued just outcomes.[21] It is this conception of a citizen which Rawls suggests should be modeled in the original position. It is to this modeling which we now turn.

---

[20] As with so much else in *Political Liberalism,* Rawls does not present a rigorous derivation of this reasonable moral psychology from the attributes of a citizen. We will not attempt to supplement what he clearly did not see as a deficiency.

[21] That citizens have a reasonable moral psychology is essential for Rawls' argument for the stability of his proposal. It is mentioned here to provide more flesh to the anorexic notion of the citizen.

Rawls now has agents appropriately constructed from which freestanding principles of political justice may be derived. These agents are the products of a democratic regime; the principles of justice (and institutional organizations) which these agents would endorse through their reasonable and rational faculties are appropriate for a constitutional democracy. But how do we conclude which state of affairs they would endorse? These agents say nothing themselves, and were we to attempt to speak for them – as we surely must - what assurance do we have that our own social station or world views, or in short, our unreasonable and irrational tendencies, would not obscure the correct response? It is in response to this problem that Rawls proposes that we attempt to divine the appropriate answer to this now pressing question by way of the original position.

We will not dwell upon the specifics of the original position and the veil of ignorance here, since we already discussed them at length during our investigation of *A Theory of Justice*. It will be enough here to discuss how Rawls thinks that his original position would overcome our own imperfect reasoning on the matter at hand.[22] In an effort to overcome our own biases, Rawls envisions not theorists directly discussing which principles of justice should be agreed to, but instead what principles of justice and discussion and, ultimately, institutions would be rationally agreed to by agents who are unaware of their situation - rational agents who are behind a veil of ignorance. This discussion cannot, of course, be fully devoid of inputs; rational agents with absolutely no awareness of their situation – including their conception of the good and any general information about the workings of the world in general – could not rationally agree to

[22] I do not mean to suggest that Rawls thinks that his solution will completely overcome the difficulty, only minimize its impact.

anything. They could not, therefore, be argued by theorists to agree to anything. These rational representatives are thus to be conceived instead as having some general information: they know that they are going to be expected to reside within the society they have created; they have access to uncontroversial information about the world and its various workings; and they have a set of interests supposed to be invaluable to any person as a citizen. Theorists, then, are to discuss what principles *these* agents would find in their best interests. It is these principles, once the theorists have agreed upon what they are, that are deemed by Rawls to be the most appropriate.

Rawls suggests that the reasonable elements of people as citizens are represented in *Political Liberalism* by the constraints of the veil of ignorance. Only reasons that can be acceptable to an agent supposing that he could turn out to be any citizen will be offered and accepted in this model, and therefore be acceptable to all. The rational elements of citizens are modeled by the uncontroversial empirical claims, and by an index of primary goods with which the theorists reason about the most rational set of principles which could be agreed to under these conditions. In such a way the citizens' conclusions about the principles and institutions most appropriate to guiding society as a fair cooperative endeavor over time, as well as the reason appropriate to public discussion of political matters, could be mimicked.[23]

II.ii Moral Consciousness and Communicative Action

Jürgen Habermas judges all social action to be either communicative or strategic. Strategic actions are ones in which an actor seeks to modify the behavior of one or more

other actors by means of economic dis/incentives, in ways that the first actor has deemed instrumentally valuable. Communicative actions are ones in which actors seek to coordinate behaviors or speech acts in such a way as to facilitate a common understanding between them. Notwithstanding the possible economic reduction of the second type to a special case of the first, Habermas treats these categories as different in kind; strategic interaction merely seeks to influence behavior, while communicative interaction seeks to internally motivate participants to accept as valid various claims. It is the active acceptance of these claims that differentiates an understanding from mere agreement or accord. Accord could be produced strategically; people could be in agreement regarding the content of an utterance merely by accident. Understanding is the result of a unanimous rational assent to various claims and *reasons*. Understanding is achieved through rational assent. While Habermas suggests that these categories exhaust the possible types of social action, he does not suppose that they are exclusive: a communicative action may have strategic elements within it, and *vice versa*.[24]

As Habermas does not recognize the possibility of individual action geared towards understanding, for him communicative action plays an invaluable role in the development and maintenance of our psychology. This becomes clear when one attends to the significance Habermas attaches to discourse. Discourse is presented as a special case of communicative action, where participants reflect on the linguistic utterances of each other in order to test various claims that have been identified as problematic. Internal discourse is a special case of discourse proper; a reflecting agent considers a case

---

[23] As Rawls notes, it is useless to derive principles of justice without also deriving an acceptable method of reasoning with which to discuss how best to institutionalize these principles. See especially *Political Liberalism*, p. 224.

from two (or more) perspectives, and judges the merit of their various reasons, eventually endorsing one and rejecting the others.

A child first develops within a culture, and internalizes a system of beliefs.[25] This system of beliefs is eventually called into question by the recognition of external pressures – primarily other people engaging the child in communicative action. Insofar as the child recognizes the force of the reasons offered by the other person, that child's belief system is then modified in such a way that it can accommodate those reasons. In such a way the child's psychology develops. Each person's beliefs and desires are maintained and refined through communicative actions throughout her life. A personality is refined when discourse results in a modification to the structure of a person's beliefs or desires through the force of reason. Even when no change to the content of one's identity results, a discourse still has the effect of re-affirming that content, and thus maintains an actor's previous personality.

In each communicative action, a person makes claims about either, the 1) objective world, 2) the social world, or 3) the subjective world. [26] That is to say that a person may be making a claim about the world, or about the norms of the speaker's (and hearer's) community, or about the speaker's own personal (internal) experiences. In engaging in communication, all speakers also claim that their statements are valid – that their statements about the world are true, their evaluative statements accord with the

---

[24] See especially Habermas 1990, p. 140, & Habermas' *Theory of Communicative Action*, Vol.1, Boston 1984, pp. 285-287.

[25] For the purposes of clearly demarcating only what is of central importance for our purposes, I will here merely mention that Habermas has a fairly developed account of the relations between each person's subjective belief system, the social world in which it develops, and the objective world. Our purposes will not be served by further elucidation, however.

[26] While these claims are recognized to be both verbal and non-verbal, we will focus only on the verbal case for the sake of readability – Habermas supposes that what may be said of the one may also be said of the other.

norms of their community (are right), and that their claims about their personal experience are honest. Each of these types of claims may be either disputed or accepted by the hearer, with agreement obtaining only when each participant endorses the same set of claims. Understanding obtains when there is agreement on the claims, as well as on the grounds upon which those claims are made.

The origin of a communicative action is a two-stage process. In the first stage a speaker offers a claim to a listener. In the second stage a hearer responds to the original claim of the speaker. They have then mutually embarked upon a communicative action. As can be gleaned from the above, each speaker has made, by implication, certain claims about the utterance that he has made, claims to the effect that it is true that X, or it is right that X, or I feel that X. Further, Habermas suggests that the speaker has, by making the claim publicly, offered to redeem his claim should this prove necessary. The hearer has accepted this offer, and in replying has offered similar assurances. In engaging in this activity, the various participants have thus created certain obligations, have accepted as binding certain norms of behavior that will in turn be the bedrock upon which all justificatory attempts may depend. These standards are inevitably endorsed and intuitively understood by any competent speaker engaged in communicative action, and cannot be denied upon pain of contradiction. They cannot be acted against upon pain of performative contradiction.

While Habermas produces only a sketch of his proposal, there is certainly enough detail to make clear his argument. He suggests, following R. Alexy's analysis, that an actor engaged in discursively redeeming normative claims must be committed *at least* to the following: allowing all people to participate; allowing assertion of any claim

whatever; allowing inquiry regarding any claim whatever; and disallowing or forbidding all forms of coercion.[27] Denying these, Habermas argues, leads to a contradiction. Indeed, all argumentative presuppositions are to be identified by showing that one who denies these claims is "caught up in performative contradictions."[28] It is these particular presuppositions, however, when combined with an understanding that belief justification (and thus normative justification) *must* be based on agreement reached by argumentation,[29] that together justify his *principle of universalization*,

> that every valid norm must fulfill the following condition: (U) *all* affected can accept the consequences and the side effects its *general* observance can be anticipated to have for the satisfaction of *everyone's* interests (and these consequences are preferred to those of know alternative possibilities for regulation).[30]

This is to say that each actor who engages in a discursive attempt to justify a claim, also implicitly acknowledges (U). While the argument is not explicitly stated, it seems obvious that it is only given each person's ability to accept the consequences of the observance of a particular norm that they may be expected to endorse the norm at issue. This leads rather directly to Habermas' principle of discourse ethics (D):

> Only those norms can claim to be valid that meet (or could meet) with the approval of all affected in their capacity as participants in a practical discourse.[31]

Anthropomorphism aside, clearly Habermas means to suggest that a norm with any claim at all to validity is one that would be the result of an actual practical discourse. Recalling that Habermas states that the validity of a claim about the social world depends on its being recognized by the community, that some form of consensus is necessary is

---

[27] R. Alexy, "Eine Theorie des practischen Diskurses," in W. Olmuller, ed., *Normenbergrundung, Normernurchsetzung,* (Panderborn, 1978)
[28] Habermas, 1990, p. 89.
[29] See esp. Habermas 1990, p. 14.

hardly unforseen. What might have been unforseen is that Habermas suggests that norms must be justified in actual discussions. Habermas suggests that any attempt to justify norms hypothetically is doomed to fail, as it is not justified in discourse. Thus Habermas distances himself from other idealized discourse theorists, such as John Rawls.[32]

While it is not my intention to here bring Habermas to task for his shortcomings, it will be helpful to present his response to critics who suggest that his proposal makes exclusive use of action geared towards understanding, and thus cannot be applied to strategic interactions. While agents may be committed to the above-mentioned principles of communicative action, it has not been argued that these principles are also intuitively understood to govern strategic interaction. Indeed, given Habermas' definition of strategic interaction, it seems unlikely that these principles could be derived from strategic interactions. As a result, the critics suppose that these principles are not applicable to strategic interactions. If this is so, surely Habermas' account is lacking, as we suppose that *if anything* is to be governed by justified norms, it is our strategic interaction.

Habermas' reply is to suggest that there is no such thing as an interaction that is *simply* strategic. This separation of strategic vs. communicative action is simply analytic – no actual agent can act without acting in his or her social structures, and therefore each action has a social element. It is this social element which entails that each action has a communicative function. If this is so, then each actor does, for each action, implicitly accept the validity of the norms of discourse. Habermas supposes that an agent cannot become detached from his lifeworld – it is too integral a part of each actor's ego system.

---

[30] Habermas, 1990, p. 65.
[31] Habermas, 1990, p. 93.

II.iii Motivating Force and Discourse Ethics

In order to develop a promising line of criticism of Rawls' *Political Liberalism* we first turn to Habermas. Habermas, unsurprisingly, decries Rawls' use of the veil of ignorance as well as the use of theory at all in an effort to bring about a meaningful discourse ethic. Seeing the entire project as an effort to either induce the reader to come to an understanding of what would be agreed to in Rawls' described situation, or as a way for each theorist to come to his or her own conclusion and then allow for a majority ruling to produce the correct normative conclusions, Habermas concludes that either interpretation is doomed to fail. If each person were to meditate over *Political Liberalism* and come to his or her own conclusion over what would be agreed to, this would not express a "common will."[33] That is to say that this could not count as internally motivating agreement. Neither would a majority vote by theorists count as an internally motivating agreement expressive of a common will. Nothing short of actual agreement as a result of actual discourse would produce the necessary results. Rawls' response is both uncharacteristically brief and weak. Rawls replies that it is his intention to have his project judged by all citizens over time, and so it is neither of the projects supposed by Habermas. It is also, therefore, not prone to the critique proposed by Habermas.

Assertion quite aside, Rawls' own interpretation of his project is still uncomfortably close to the second of Habermas' interpretations. Everyone judging a project over time still seems most naturally interpreted as each citizen reading *Political Liberalism* and then *afterwards* taking what he or she wishes from it. Nevertheless,

---

[32] Habermas, 1990, p. 66, fn.
[33] Habermas, 1990, p. 67.

whatever interpretation one cares to give Rawls' project, it still clear that it cannot be interpreted as an actual conversation. The sheer number of necessary participants alone disqualifies it from consideration as such. If it is not the result of an actual conversation, then the proposed norm has not yet met Habermas' conditions for knowledge of its validity.[34]

Habermas' point is quite well taken: without an actual agreement, no amount of supposition that there may be agreement on any particular norm produces knowledge about its worth. Moreover, there are concerns about the motivating force of hypothetical agreements that are distinct from questions about the accuracy of our foreknowledge of the content of any actual agreement. For in so far as a proposed norm is not an actual expression of a general will verified in conversation, it fails to have the internal motivation necessary for a truly binding set of norms to result. Recalling Ronald Dworkin's famous critique of *A Theory of Justice*, we suggest that the hypothetical conclusion of a hypothetical discussion fails to produce in actual people a motivation for recognizing the force of the proposed norm.[35]

Also following Dworkin, we recognize that it would give people a reason to accept a proposed norm if it were to be shown to these people that they would, *right now*, accept those conclusions. For this to be the case, interestingly, an actual discourse may be said to occur. But could Rawls' project be reasonably supposed to provide such arguments? Rawls suggests that we all try to adopt the position outlined by the veil of ignorance, and attempt to reason from within this blind as citizens. Each suggestion is a

---

[34] It should be recalled that, for Habermas, the norm could be said to be valid if it *would* be the result of a discourse. But I take Habermas to be objecting that we cannot have knowledge of the validity of this norm without it being agreed to in an actual discourse situation.

significant idealization. In an actual discussion, each person would reason with his or her full information, as himself or herself. Any idealization of the discussion situation must, in order for us to suspect that the conclusion would be one that we would accept, be supposed to accurately reflect what we already believe, and the degree to which we believe it. But this would render the idealization entirely unmotivated. Any other idealization can only be reasonably supposed to distance the conclusions from the ones that we would ourselves come to. It would be only by way of happenstance that the modifications to the discourse situation resulted in identical conclusions to what would result in an actual discourse situation. And so we conclude that these idealizations contaminate the results; the grander the idealization, the further we may expect the conclusions to be from peoples' actual conclusions. This separation, in turn, undermines the practical force of the argument, the force we earlier found to be necessary for a justification of any set of norms.

Habermas' proposal faces no such difficulty. But it overcomes this hurdle by running the risk of eliminating the use for theoretical apparatus in the justification of norms. The only possible place for theoretical devices would be to exactly mimic the argumentative processes that actually occur, and to present the conclusions derived for the consideration of the participants who have been so mimicked. This makes the value of a theoretical discourse ethic questionable.[36] Habermas, of course, entirely agrees.

Something that may be more troubling to Habermas is the suggestion, mentioned above, that strategic interaction is not regulated by discourse ethics. Notwithstanding his

---

[35] See chapter six of Ronald Dworkin's *Taking Rights Seriously*, Harvard University Press (Cambridge, Massachusetts) 1978. Esp. pp. 150-154.

[36] Instrumental contractarians, of course, face an obviously analogous problem, but it is best dealt with below.

assertion that there is no such thing as a simply strategic interaction, but only strategic actions that have communicative elements, and which are thereby governed by justified norms, one could simply deny this fact. Habermas leaves his assertion entirely undefended.

Moreover he provides no reason to suppose that the existence of some communicative elements in any given strategic interaction ought to lead one to conclude that the communicative elements ought to entirely govern this mixed-motive action. It seems more reasonable to suggest instead that the strategic elements of any given action are still not bound by justified norms, while still allowing that the communicative elements of any given action are so bound. Or that to the degree that a mixed-motive action is a communicative action, it is *to that degree* governed by his justified norms. Given that Habermas supposes that the elements can be separately identified, surely he must give reasons beyond mere assertion for the governance of strategic actions by communicatively justified norms before we may be expected to find his account compelling.

Allowing that they may not be separated, however, one must still suppose that the strategic elements of any strategic action together compose the greater part of that whole – indeed this seems analytic. Given this, it may be supposed that strategic considerations would ultimately lead to that action. There may be some communicative concerns involved in the deliberation, but the strategic concerns would be decisive, as they are the greater part.

If this analytic argument is not compelling, it is to the nature of people to which we finally turn. Any justification of a moral and political system must display practical

force, it must motivate the adoption of that system. This force must not be slight, nor easily overcome. Perhaps pessimistically, we suppose that communicative concerns are not the largest motivating force that pulses within the human brain – we suggest instead that economic concerns are the larger part of human motivation. Any who deny this are simply ignorant of the empirical evidence. Take, for instance, the interest in academia, deliberation, debate, discussion, etc., and pit it against the interest observed in the world of business. Further, look at the need for the enforcement of contracts – the need to force people to honor their word. Lastly, look to the prevalence of hypocrisy both now and in ages past. There is significant evidence to suggest that people prefer to pursue economic advantage even if it means forsaking their reasoned principles. The motivation derived from the acceptance of good reasons is by far the smaller part of humanity. If we are to present a justification of moral and political systems, it must be by way of appealing to the economically rational side of humanity. This does not discount the appeal to reason at all; this economic argument must be presented to the intellect. But the grounds upon which this argument is to be constructed must be economic if it is to be reasonably assured of convincing that intellect at the time of the presentation.

III. Sociobiological Justifications

We now turn to theories that attempt to explain the creation of moral or political institutions by way of evolutionary arguments. Examining natural or constructed agents in their environments, these theories explore the development of moral behaviors and social structures. Attempting to justify morality by explaining the evolutionary circumstances through which it was created seems an initially plausible route. If an

argument can be forwarded which explains how certain moral behaviors evolved, it will presumably give reasons for the adoption of moral behaviors. Although attempts to justify moral constraint by way of evolutionary apparatus are many, we will examine in any detail only Peter Danielson's *Artificial Morality*.[37] Danielson himself dismisses the vast bulk of evolutionary arguments summarily, suggesting that they either fail to justify moral constraint, or they fail to justify moral behaviors as constraints. We will briefly explain this further before proceeding to examine Danielson's positive thesis.

Danielson first presents a general argument designed to exclude any direct appeal to sociobiology. Sociobiological appeals proper tend to explain only that certain behaviors are exhibited in certain organisms – behaviors that we consider moral. They do not, and cannot, claim that an organism finds these behaviors useful to adopt. Indeed, given the tendency to explain evolution as an unconscious method of selection of particular traits, it would be quite surprising if sociobiology proper ever suggested that these traits were adopted consciously. Sociobiologists can explain particular benefits derived from these behaviors, but cannot claim that these benefits are the reason why these behaviors were adopted. The reasons are only discoverable by recourse to internal processes. Sociobiology, however, does not lend itself to the examination of the conscious psychological processes of its objects of study. Since this is so, it is ill equipped to explain the rationality of the adoption of certain constraints to their behavior.

Some game-theoretic sociobiological attempts to examine the justification of moral behavior suggest that, given that agents who live together in an environment over time are likely to come into contact with each other several times, and given their preferences, it is rational to exhibit some constraint in some strategic situations. To more

---

[37] Peter Danielson, *Artificial Morality* Routeledge (New York; New York), 1992.

successfully interact with the other agent in the future, one ought to modify one's behavior now. Cheating on an agent the first time you interact is not all that smart, if she will thereafter cheat you on every subsequent interaction, or not interact with you again. It would have been better to cooperate all along. Although this line of inquiry avoids the problems associated with sociobiology proper, Danielson nevertheless discounts this possible route for the normative theorist. It will be recalled that Danielson suggests that moral behavior must involve constraint. Showing that it is straightforwardly rational to play these games a certain way shows that no constraint is necessary. As such, these so-called justifications hold no promise. And so Danielson is committed to presenting a justification of moral behavior that constrains our otherwise rational behavior that is straightforwardly psychological. It is that project to which we now turn.

III.i Artificial Morality

Wishing to present a fundamental justification for morality, Danielson argues that moral agents are substantively rational. Wishing also to take issue with the received theory of rational choice, he proposes a heterodox theory of rational action, one with which he provides arguments designed to suggest that agents following his theory of indirect rational choice fare better than agents committed to the received theory of rational choice. Danielson's challenge to the received theory of rational choice is not unmotivated; he supposes that the traditional account cannot be successfully made use of to produce a fundamental justification of morality. Suggesting that the structure of rational choice theory (RCT) makes providing a fundamental justification of morality impossible, an alternative account of rational choice must be provided.

Upon providing such an alternative, Danielson proposes to test it against the received theory, as well as to test amoral versus various moral agents, in a simulated arena. Danielson runs computer simulations in which various agents are constructed to interact strategically with each other, to see which would fare best.[38] Since the amoral agent is the agent Danielson suggests best represents the recommendations of orthodox RCT, the results of his computer simulation could provide evidence for both the thesis that morality is rational, and the thesis that his heterodox RCT better achieves the goal of instrumental rationality. This presupposes, of course, that a neutral criterion for instrumental success could be determined. Danielson suggests that an agent whose substantive interests are better served by his patterns of strategic interaction than some other agent, when interacting in mixed-motive games, is to be identified as the more rational one. Similarly, a theory of rational choice that recommends courses of action which (when followed) better serve an agent's substantive interests than some other such theory, is to be identified as the more rational strategy. But this description is heavily laden with technical terms, ones which demand a more detailed analysis. Such an analysis is best provided within a more detailed accounting of Danielson's project.

As mentioned above, Danielson borrows the concept of fundamental justification from Robert Nozick's *Anarchy, State and Utopia.*[39]A fundamental justification is first defined as "a justification of a realm that does not appeal to any of the concepts in that realm."[40] Danielson latter adds that a fundamental justification of morality must provide sufficient motivation for following the dictates of morality; anything less would produce

---

[38] Danielson takes quite a bit of time dwelling upon the procedural problems of constructing this computer environment, and the particular agents. But it is not our intention to critique his implementation, but more importantly, his very project. So we will completely pass over his procedural discussions, including his descriptions of the various agents constructed.

an explanation of moral theory, at best. To provide such motivation, instrumental rationality seems the obvious autonomous realm to which one should turn. To make use of instrumental rationality to provide a fundamental justification of morality, one must show that morality, as an effective constraint upon agents' choices, *is* rational; that is, that agents who are moral perform better in social situations than their amoral counterparts. Danielson proposes, following the categorization of issues by Alan Gibbard, to focus on the problem of fidelity. [41] This is the problem of providing an argument for abiding by standards of interaction agreed upon by two agents. As we are now aware, this proposal does not require that all social situations must be examined. Many situations are ones in which constraint is not necessary; one's unconstrained choices lead to the best solution for everyone involved. Iterated interactions in which not cheating on each particular action is warranted given future expected behavior, as previously mentioned, are of this type, as are coordination games. Morality is not required in these situations. Morality is also not possible in some situations. In zero-sum games, where one person's gain is another person's loss, there is typically no 'moral' course of action. There could be no agreed upon standard for interaction; whatever pair of actions is chosen, one person will be the 'loser'. It is only in mixed motive games that morality becomes possibly justifiable. They involve situations where it makes sense to propose rules dictating particular choices, and also makes sense to fail to comply with these rules once they are agreed upon.

---

[39] Robert Nozick, *Anarchy State and Utopia*, New York; Basic Books, Inc., pp. 6-9.
[40] *Artificial Morality*, p. 20.
[41] Alan Gibbard, "Norms, discussion, and ritual: evolutionary puzzles", *Ethics*, 100: pp. 787-802. Danielson forgoes, in an effort to produce a manageable project, accounts of justice (the standards to set) and allegiance (fidelity for large groups.)

Danielson supposes that the Prisoner's Dilemma (PD) is the game that accurately models the problem of fidelity. In a PD, each agent prefers to defect instead of cooperate no matter what the opponent does, but if both defect, they do worse than if they had both cooperated. Knowing that this is the situation they are in, each agent realizes that they will both defect, all other things being equal, and that they could do better by both cooperating – there is room for an improved outcome to result given agreed upon courses of action. But even if the agreement is struck, there is no reason to expect that either agent will adhere to this agreement – they would each do better by defecting. To provide a fundamental justification of morality, one must provide an argument which proves that moral agents (agents who follow the dictates of their moral rules) fare better than amoral agents when playing PDs. Danielson supposes that this is an impossible task, given the received rational choice theory, and therefore is motivated to provide an alternative. If he cannot, he must conclude that his project is doomed to fail.

Danielson has good reason for pessimism given his understanding of rational choice theory, game theory and morality. Danielson understands RCT as a theory of deliberation given idealized agents. Agents are supposed to be equally rational, have access to each others' subjective preference ranking over outcomes, and have similar information about the situation at hand. These agents perform calculations in parametric and strategic choice situations by which they determine which course of action may be expected to best satisfy their preferences, and they act so as to attempt to maximize on their preferences given that calculation.[42] It is by reference to the choices presented, as well as to the patterns of preferences, that the games being played are identified.

---

[42] Given that morality is exhibited in strategic choice situations, it is only such situations that Danielson discusses; we will follow his example.

In a PD, each rational agent has two choices: to cooperate with his or her partner in the game, or to defect. Each agent prefers to defect given that his or her partner also defects, and also prefers to defect given that his or her partner cooperates. Both agents also prefer a joint cooperative outcome to the joint defection outcome. Unfortunately, given rational choice theory, these agents are forever doomed to end up both defecting. Given that each agent acts in such a way as to maximize his or her preference satisfaction and given that a PD is, in part, *defined* as a game in which each agent's expected preference satisfaction would be maximized if he or she defected, rational choice theoretic agents thus defect. In any situation where the calculation suggested cooperation, they would not be playing a PD; cooperation would have to be preferred to defection. Agents do not have the ability to constrain their behavior, and morality is irrational by definition.

Since rational choice theory defines agents in this minimal way, they are also symmetrical in a particular way; given the same preferences, options and the same understanding of a particular choice situation, each agent would choose the same action. Danielson suggests that this too is unacceptable. Symmetrical agents are too simple to allow for a robust solution to the problem of fidelity. It is by allowing the possibility that agents could act differently from other agents in mixed motive games that Danielson, following Gauthier, argues for the rationality of fidelity.[43] By allowing agents to choose to cooperate with other agents willing to do the same, and not cooperating with agents who are not willing to also do so, Danielson and Gauthier both argue for the rationality of

[43] Gauthier's most familiar account being in *Moral by Agreement*, Oxford University Press, 1986. Chapter five of this thesis presents a detailed account.

acting on agreements.[44] Symmetrical agents simplify the situation too drastically, as there would be no diversity of agents to treat differently.

Danielson's alternative account of rational choice theory, then, must allow for a gap between chosen outcomes and an agent's preferences. For in no other way can morality, understood as a constraint on otherwise preferred courses of action, possibly be rational. This also opens the door for the possibility of non-symmetrical agents. Two agents might have the same choices open to them, face the same type of agent, be playing the same game, and nevertheless choose different courses of action. This might be accomplished in one of two ways, while remaining roughly within the domain of instrumental rationality and game theory, as standardly understood. One could allow preferences to determine the choice of actions, but not to determine the game being played, or one might allow that preference determines the game being played, but not the agent's choice of action. Danielson chooses the first of these possibilities, allowing that preference will determine the course of actions, while an agent's interests will determine the game being played.[45] These agents play games that are constructed from each agent's objective interests. While unclear about the objects of these interests, he does suggest at least that they are things that an agent *ought* to desire; given his close ties to sociobiology, we could perhaps identify these interests with things like food, or effective shelter without distorting his intent.

Now that the alternative account of rational agents' psychology is in place, and recalling that Danielson wishes to allow for asymmetrical agents in the population, we may begin to examine what is to count as a rational. First we must recognize that a type

---

[44] Although they do so in different ways, as will be clear when comparing this section and chapter five.
[45] We return to the second possibility in the critique below.

of agent is now to be identified by the preference structure he or she posses (hereafter referred to as its 'disposition'). Danielson identifies an agent as rational (or not) by comparing how well its interests are satisfied when playing PDs in mixed populations of agents to the level of satisfaction attained by the other agents. It is, in fact, quite a bit more complicated than this, given that he wishes to consider the rationality of several different types of agents in various combinations.

Danielson proposes a pair-wise testing procedure. Disposition A is more rational than disposition B if and only if it can invade a rational extension of an initial homogeneous B population and B cannot invade any rational extension of an initial homogeneous A population. A *rational extension* of a population consists of that original population, as well as some agents possessing different dispositions than any of that original population, who fare at least as well as some member of the original population. Disposition A *invades* a rational extension of disposition B by better looking after its interests when playing PDs with that population. With this test in hand Danielson proceeds to develop an account of the rationality of particular moral dispositions by programming various agents to interact with each other in PDs and attempting to provide such proofs as he can that a moral disposition is more rational than various others. Danielson takes a special interest in the disposition whose interests track its preferences, thereby effectively closing the gap between preference and game description, and embodying simultaneously the amoral agent and the standard account of rational choice theory.

III.ii Indirect Choices and Justifications

We have come a long way from sociobiology proper to Danielson's project, and it may be worthwhile recalling where that road originated. External investigations of moral constraint merely explain that particular organisms act in particular ways, and cannot produce a justification for acting in a moral fashion. Therefore an internal examination of the agents must ensue. Any internal examination that provides evidence that particular (supposedly) moral behaviors are, after all, rational must also be discounted, as they cannot explain moral behaviors as constraints on rational behavior. Therefore any examination of iterated games, or any other argument that proceeds from a traditional rational choice theory must also be discounted. In order for a justification to involve the notion of constraining rational behavior, and therefore be a possible justification of morality, Danielson supposes that a theory of action must be adopted which opens a gap between the preferences of an individual, and the chosen outcome.

We turn first to Danielson's choice of separation between chosen outcomes and preferences. Preferences determine the course of actions, while an agent's (objective) interests define the choice situation. Unfortunately, while his 'constrained' agents may then be seen to be rational in his favored sense, this is not a sense that agents themselves care about. They are shown to do better on a scale that they do not prefer to do better on. Since this is so, while they may be seen to be more successful on what can be described as an evolutionary scale, this cannot generally motivate people to act so as to maximize on that scale. One would have to rely upon an assumption that agents care about their objective interests. This is obviously entirely unsatisfactory, for it closes the gap that it has been found to be necessary to open if moral action is to be understood as a constraint

on rational behavior. Agents constructed in this way, and arguments presented with these types of agents in mind, cannot provide the internal motivating force necessary for a justification to be the end result.

To illustrate this point, imagine that a population that interacted regularly could either choose to help each other harvest their corn crops, or not. Further, while no one could harvest all of his or her crops without help, with the help of only one other person, each person could have all of his or her crops harvested. Assuming that arrangements could be made such that all of each crop was fully harvested (with suitable aid), would these agents conclude that the smart thing to do was cooperate? Would pointing out that such cooperation would maximize the amount of corn harvested convince these agents to so act? Not at all, unless one supposed that the agents *also* cared about maximizing their corn yield. Maximizing returns on some objective scale does not motivate agents to so act. Any attempted justification would lack the practical force necessary for such a justification to be judged successful. Buttressing this argument with claims that people also ought to care about the maximization of whatever objective scale was made use would only provide sufficient support if it ultimately relied on the maximization of expected utility. This in turn entirely undercuts the reliance on Danielson's proposed argumentative structure, whose point was to not so rely.

Turning now to the other alternative, one might suggest that an agent's objective interests define her choice of action, while an agent's preferences determine the choice situation. But assuming that agents could not act upon their preferences renders the entire attempt useless. Agents could not change their choice of actions even if they wanted to. They would be frustrated automatons, doomed to forever act in ways that they

do not wish; and no amount of wishing that it were otherwise would produce any external effect. Any effort to save this theory by the introduction of a psychological realm that one cares about, but that is not the realm of preference, seems at best to be a linguistic game, or a just so story.[46]

The separation of action, situation and preference thus renders the project impossible. All that remains, then, is to presume that either I) the effort to provide a fundamental justification of morality derived from rationality is impossible, or II) that Danielson's understanding of rational choice theory or morality as constraint is flawed in some way. Unsurprisingly, we choose the latter. Danielson's suggestion that standard rational choice theory is incapable of accommodating moral theory because it must constrain maximizing behavior is incorrect. The evidence that leads one to suppose that constraint is a necessary part of moral action only does not support this contention. The intuition that constraint is a necessary part of moral activity has strong opposing intuitions that it cannot easily overcome. We suggest that morality be understood in a different sense than it is currently, one in which constrained activity is not a necessary condition of moral action, but rather some evidence that one is trying to become a moral person. This will be discussed at greater length in chapter 5.

IV Conclusion

And so we conclude our investigation of alternative attempts to justify moral and political systems. Instrumental contractarianism is what remains of this particular cluster of theories. The instrumental contractarian is one who relies exclusively on the

[46] This critique assumes that one's preferences do not contribute to the form of one's objective interests. Any model developed that assumes such a relation would bear such strong similarities to Gauthier's

preferences of each economic agent to produce an identification of, and motivation for adhering to, the dictates of some set of moral and political systems. Of course this is, as of yet, a frustratingly general characterization. To develop the details of an acceptable contractarian theory is, in part, the focus of our next chapter. When doing so, we will be well advised to keep in the front of our minds the lessons gleaned from the above explorations.

While we must produce theoretic agents if theory is to result, we must minimize the idealizations. We must take care to minimize the distortions between the theoretical agents and the actual agents to whom our results are to be said to apply. Only then may we expect our results to produce compelling reasons for actual agents to endorse them, and therefore be worthy of the title 'justification'. Of course it stands to reason that we must justify delving into the realm of theory – to deny that the conclusion of real world situations *is* what is justified given economic interests. Since the justification of the particular characterizations of the agents, as well as the particular argumentative structure being made use of, may not rely on our moral or political intuitions, other grounds must be presented. Failure to do so would render our argument circular, and not particularly interesting. And finally, we must not separate action from preference; to do so is to doom the project for lack of motivation. We must nevertheless maintain this connection while making sense of the idea that morality is a constraint upon action. This is, it must be noted, different than justifying the theoretical apparatus by way of appealing to our basic intuitions about the matter. We will argue separately for the theoretical form of the project, while showing that this theory will still produce a result consistent with our basic

---

account of rational choice theory that the reader should refer to chapter five of this thesis for a sustained critique of this approach.

understanding of the subject matter. It is with these lessons and goals in hand that we proceed to the development of the contractarian enterprise and the rendering of the most compelling of these projects into mathematical form.

Chapter 2
~ Tractable Contractarianism~

We are concerned with fundamental justifications. Examining the concept of a fundamental justification for moral and political theory, we conclude that it must not rest on biases and that its recommendations must have practical force. A fundamental justification of a moral theory should provide reasons to act in order to rise above the status of a mere explanation. Taking seriously the necessity for the justification of a moral or political theory to have practical force leads us to examine peoples' psychology. We observe that the mere public espousal of the validity of certain norms of interaction fails to provide the proclaimant sufficient motive to act in conformity with them; rational calculation often overcomes that which one has said one ought to do. Discourse ethics are not found sufficiently motivating. We turn then, naturally enough, from public espousal to that which is seen to override it. In such a way we come to conclude that a fundamental justification of moral and political principles ought to rely on the economic interests of people. Finding fault with the alternative accounts of rational choice theory, we make use of an orthodox conception of rational choice with which to examine economic interests; an action is rational when it is thought, by that agent, to be the response most likely to maximize the satisfaction of his or her desires in a given situation.

Those conversant in rational choice theory will likely be raising eyebrows to almost comedic heights at this claim. Having only dwelt upon Danielson's alternative account of rational choice theory, we can hardly be said to have surveyed the field. Further, we have no intention of surveying the entire field. Such a project is clearly a work in and of itself. But it is clearly one benefit of endorsing the orthodoxy that one need not feel compelled to assume the burden of proof. We will, however, deal with

Gauthier's heterodox solution, which involves the introduction of rational dispositions, in chapter 5.

I: Instrumental Contractarianism

We take instrumental contractarians to be those committed to providing a rational basis for moral and political constraint.[47] Contractarian accounts are generally understood as either seeking to convince each person that certain rules of interaction are, all things considered, the rules to adopt, or as presenting a set of rules that everyone would (should) rationally adopt. These two projects may seem indistinguishable at first, but there is a subtle, and crucial, difference. Only the former can account for the internally motivating force of the set of endorsed constraints. Merely identifying a set of principles to govern peoples' interactions is not sufficient. Such identification can only be the first part of the contractarian enterprise. Contractarians must always ultimately aim towards presenting this set of identified principles to the people to whom it is to apply. The identification means almost nothing without the adoption. Anything less falls prey to a critique that takes both Habermas and Dworkin to heart: theorists cannot conclude that the hypothetical motivational force of the arguments for their particular principles thereby actually motivates people to fit those principles. If a theory cannot in principle be presented for consideration, then it cannot be supposed to have a chance at being endorsed. If a contractarian account cannot in principle be endorsed, then it cannot be argued that it is compelling. If it cannot be argued that it is compelling, then the contractarian account must be found wanting.

---

[47] Given that the remainder of this work deals exclusively with instrumental contractarianism, we will hereafter use drop the word 'instrumental' when referring to this theory.

But this naturally enough raises the question alluded to in chapter one: why, then, do we need theory at all? Why aren't we all doomed to answer these questions on the political stage, and not in the academic's armchair? The answer, I suggest lies in our recognition that people reason imperfectly and slowly. Each person's preference set is inconsistent and incomplete, each person's set of beliefs is inconsistent and incomplete, we reason abductively, and we rely on heuristics that lead us astray in crucial cases. This is to say nothing of the fact that we don't, in general, wish to spend all of our lives bargaining with our neighbors about the proper form, function, and content of the rules governing our behaviors. Theory, in short, may afford us an answer to which we would not otherwise have access. Our technology might provide for us resources of which we would otherwise not be able to make use.

And so the contractarian attempts to identify a set of rules that people would rationally endorse once brought to their attention. These rules are to govern how one may interact with another, and so one must understand what is to count as an interaction. An interaction may most broadly be defined as any two or more persons engaged in activities in such a way that the actions of at least one of these persons affects at least one other person, and that each person in the interaction is either affected or affecting.[48] A *legitimate* action is one approved of by a set of rules. That is to say that for a given set of rules R_j, an action is legitimate according to R_j if it is permitted, required or not forbidden by R_j.[49] These rules must also delineate what is to count as property; for it is impossible

[48] *Contra* the common usage, this definition allows that a pilot who drops a bomb interacts with every inhabitant of a particular city, as has been brought to my attention by David DeVidi. It is also worth pointing out that on this understanding, what goes on in my bedroom can be an interaction involving people who are not physically present in it.

[49] We will hereafter only be considering complete sets of rules, where each action is required, permitted, or forbidden. Thus 'not forbidden' becomes equivalent to being either required or permitted, though this is not true in general (i.e., of non-complete rule sets).

to delineate what may be done without a further delineation of what may be made use of. One's property is to be defined as that of which one may legitimately make use without the permission of any other person, and of which no one else may make use without permission.[50] The contractarian must identify a particular set of rules that has the greatest chance of being accepted by people in place of their moral intuitions. That property delineation in particular is a function of the set of rules has important consequences latter on.

We submit that the version of the contractarian project most likely to succeed is the account which makes use of an accurate identification of each person's preferences over finitely many outcomes, and beliefs regarding the likelihood of all combinations of these outcomes obtaining, given each proposed set of rules governing behavior. With this in hand the contractarian rank orders the set of all possible rules of constraint for each agent. Each person's ranking of the set of each of the possible rules of constraint is summed with each other person's, and a set of rules that receives maximal support is identified as the set to bring to the people's attention. If such a project is successfully undertaken, its conclusions would have the best chance of gaining people's allegiance. Attempting to render a maximally compelling system of rules governing actions, we attach its identification as strongly as we can to peoples' beliefs and desires, which are recognized to be motivating.

It is obvious that contractarians do not, by and large, attempt any such enterprise. At best a sub-set of people's preferences and beliefs are taken as salient. For instance contractarians have excluded tuistic preferences (Gauthier), and have only made use of preferences that most of us have, and have quite strongly (Hobbes). They have by and

[50] Modified in the obvious way, this applies to common property as well as group ownership.

large assumed, in some sense or other, the equal worth of their participants, and have made use of the state of nature as a background against which to assess their proposed constraints. Some of these assumptions only weaken the contractarian account that appeals to them, some others are, perhaps, excusable simplifications employed in an effort to simplify the contractarian's computational burden, and, lastly, some are necessary in order to achieve a meaningful result at all. To distinguish into which category each of these assumptions fall it is necessary to consider each at some length. As we must first see what the more complex project involves before exploring which simplifications are excusable, we here focus on the assumptions that unnecessarily weaken the contractarian account. And so we turn to a discussion of preference sets, belief sets and the set of all rules of interaction. With these concepts in hand, we turn to the function with which we identify the proper set of rules of interaction. Along the way it will become clear why the contractarian is to insist on the assumption of equality and why she is not to rely on the state of nature.

I.i: Preferences

Two considerations will occupy us here: the preferences that are to be allowed as inputs, and the formal constraints to which we demand that each set of preferences conform. Given the nature of our enterprise, it may be thought that there is a simple response to any question regarding which preferences are to be allowed as inputs: all preferences are to be included. Anything less than total inclusion renders the calculation less motivating than it would otherwise be to the agent whose preference has been excluded. However, there has been, as of late, much ado about tuistic preferences – preferences that make reference to the utility function of another. David Gauthier has

insisted upon the exclusion of tuistic preferences for a variety of reasons, and Susan Dimock has supplemented the reasons supporting this insistence.[51] This insistence has been hotly contested in the literature, and the matter is by no means settled. We will here briefly examine most of the reasons offered for the assumption of non-tuism, and find them lacking. Whatever the merits of this assumption for particular contractarian efforts, or economic endeavors, it can play no role here. We will then move on to consider the formal conditions demanded of the sets of preferences to be made use of.

I.i.a: Tuistic Preferences

David Gauthier makes reference to both the assumption of non-tuism and the assumption of mutual unconcern. The former is the assumption that one does not take an interest in the interests of the agents with whom one is engaging in an economic interaction. The latter assumption entails that agents are conceived of as having no interests in the interests of any other person. The distinction between the two disappears when one is engaging in an all-encompassing contractarian effort; all agents are involved in this economic endeavor.[52]

Historically, non-tuism originates in Philip Henry Wicksteed's *The Common Sense of Political Economy*, and is insisted upon as an analytic consequence of his delineation of questions of political economy.[53] Insisting that all choice is rational choice, he suggests that political economy is that subset of rational choice situations in which one engages in trade with another, and does not consider that other person's

---

[51] David Gauthier's *Morals by Agreement*, pp. 11, 85-90, 100-102; Susan Dimock, "Defending Non-Tuism", *Canadian Journal of Philosophy*, 29:2, 1999, pp. 251-274.

[52] It is common to suppose that children are not included in the contractarian choice situation. So conceived, then, the distinction does not entirely disappear, but is greatly reduced in significance.

[53] Philip Henry Wicksteed, *The Common Sense of Political Economy* (2 vols.), Lionel Robbins, ed., Augustus M. Kelly (Publisher), New York, 1967.

satisfaction while determining whether to (or at what price to) engage in a transaction.[54] Therefore the degree to which one is concerned with another person's preferences is the degree to which the transaction is not one of political economy. It is, to that degree, perhaps a matter of familial economic choice (who gets dessert, and how much of it), but not a matter of political economy. This cannot be thought sufficient to exclude tuistic preferences from our arena, the arena of general moral philosophy; we move straight-away to considerations that have been thought to do so.[55]

Rawls adopts the assumption of mutual unconcern in order to ensure that the outcome of his contractualist enterprise is fair. We have already suggested why this consideration is out of court; any appeal to moral considerations renders the project a non-starter. Fundamental justifications of morality must make use of no moral considerations when constructing the theoretical apparatus by which the result is obtained. This same consideration excludes appeals by some feminists who might suggest that women will be materially worse off should tuistic preferences be included in the bargain because of their being culturally constituted so that they care more for other people's satisfactions than men.[56] Notwithstanding the fact that it is impossible to equate 'materially worse off' with anything objectionable to the proponent of instrumental rationality, who cares about preference satisfaction, this is also clearly proposing that a moral presupposition (that women ought not to be disadvantaged) be allowed to shape the contractarian enterprise. As such, this line of reasoning must be rejected. Similarly,

---

[54] Ibid., see esp. pp. 169-174.
[55] We will not examine the use of this assumption for the purposes of simplification. Given the nature of our endeavor, such a consideration is not yet germane.
[56] Gauthier makes reference to this defense on page 11 of *Morals by Agreement.*

suggestions that it be made use of to avoid double counting the preference of some persons ought also to be dismissed.

The assumption of non-tuism has also been insisted upon in order to ensure that the contractarian calculation is of a fundamental character. Noting that some recognizably moral rules ("Care about others!") may be construed as tuistic in nature, it may be suggested that, as these moral rules may not be made use of, preferences that seem to obey those rules must also be excluded from the consideration. However, as Susan Dimock has observed, not all positive tuistic preferences are moral preferences. Even should we dismiss this sensible observation, the inference that a preference that recommends some action that would also be recommended by morality is therefore moral is surely absurd![57] A preference may be consistent with a moral rule without being motivated by it. Even if one were to deny this, the line of argument misunderstands what it is that contractarians must accomplish to avoid begging the question. As was made clear in the previous chapter, those seeking a fundamental justification must rely on no moral considerations whatever when delineating their project. These tuistic preferences are not, obviously, being made use of to shape the enterprise. Notwithstanding that their inclusion may lead to the adoption of some set of moral rules over some other set, they are not included in order to do so, and so cannot be excluded over concerns of question-begging.

Non-tuism has also been insisted upon as a way to ensure that morality holds in the hard cases – cases where no one has any positive fellow feelings for any other

---

[57] Although Dimock does not suppose that this is so. She suggests that "this argument provides a limited restriction on the admissible preference set with which contract theory can operate [.]" See Susan Dimock, "Defending Non-Tuism", *Canadian Journal of Philosophy*, 29:2, 1999, esp. p. 257. While supposing that

person.[58] But Christopher Morris has pointed out that this is not the hard case for morality. The hard case would be where people actually had negative fellow feelings for every other person.[59] This is substantially different from the assumption of non-tuism, and is not a particularly interesting observation in any event. There are obviously possible scenarios in which people are so constituted that rational cooperation or constraint is impossible. Hume's circumstances of justice are never far from the ethical theorist's mind. And while delineating the circumstances more precisely is an obviously interesting project in and of itself, it is not the project with which we are concerned. We here wish to present as compelling a version of contractarianism as possible, and counter-factual constructions of agents' motivations will not, in general, help us with this project.

At least three economic considerations have been thought to motivate the exclusion of tuistic preferences. David Gauthier notes that the exclusion of externalities, given perfect information and unrestricted activity, is sufficient (but not necessary) to ensure that an equilibrium choice is also an optimum choice.[60] The inclusion of tuistic preferences allows for externalities, and since it is important that the equilibrium choice also be optimal, the exclusion of tuistic preferences is therefore justified. Christopher Morris observes that the perfect market, which assumes a lack of externalities, thereby ensures that the resulting equilibrium is not such that some sub-set of the agents engaged

---

this line of reasoning does not justify the adoption of the assumption of non-tuism, she supposes that this line of reasoning does justify excluding preferences that have a 'moral character'.

[58] See Gauthier's *Morals by Agreement*, pp. 100.

[59] See Christopher Morris' "The Relation Between Self-Interest and Justice in Contractarian Ethics", as found in *The New Social Contract*, E. Paul, F. Miller, J. Paul, and J. Ahrens, eds. Basil Blackwell, 1988, pp. 119-153.

[60] Morals by Agreement, pp. 87-90. "Optimal" is here understood in the Paretian sense; that no one could be made better off without some other person being made worse off. An externality is any result of a bargain between agents that affects any other agent not a party to this bargain. These other agents may be beneficially affected (by the services of a lighthouse, for example) or adversely affected (by the effects of inhaling smog from a factory.)

in this original market transaction could do better by trading in their own market environment which excludes the rest of the population. Tuistic preferences allow for the possibility of externalities; having a tuistic preference for an agent engaged in a particular transaction that excludes your participation, you are nevertheless affected by its outcome. Taking this observation and running with it, one might suggest that, owing to our intuition that morality ought to include everyone, we ought to adopt the assumption of non-tuism to, in part, ensure that this is so. Lastly, Susan Dimock suggests that positive tuistic preferences ought to be excluded from rational bargaining over cooperative surpluses due to the free-riding encouraged by not insisting on such a general policy.

The suggestion that results from Christopher Morris' observation that this is a necessary assumption to make in order to ensure a system of rules universal in application can be dealt with in short order. Any credibility ascribed to that line of reasoning in general depends upon our intuition that a single set of moral rules must be universal in application.[61] Without the moral motivation, there seems no other reason to require that contractarian accounts include this assumption. Moral intuitions being inapplicable at this stage of the inquiry, we proceed to an examination of the remaining two economic considerations.

That equilibrium choice coincide with optimal choice is important for Gauthier's proof that the perfectly competitive markets are *morally free zones* – that morality as a constraint upon behaviors in a free market would be unnecessary. This does not demonstrate its importance for the contractarian calculation here considered. We are

[61] This, it should be clarified, is only the case due to the particular nature of the enterprise in which we are engaged. As Jan Narveson has suggested, it could be that the universal applicability of a contractarian account is definitive of the enterprise. One might be exploring whether it is possible for a contractarian account to be completely inclusive in scope, for example.

concerned here with presenting as compelling a case as possible to individuals for the adoption of a set of moral preferences. Individuals do not, in general, care about whether the resulting set of moral injunctions is Pareto-optimal. By way of illustration, imagine that the calculations have been run, and that, all things considered, the group of agents picked set R of moral rules. Suppose that this set is not Pareto-optimal. Further suppose that person X may be made better off and no one worse off by adopting instead set P. Given the existence of Prisoners' Dilemmas (PD), it cannot in general be supposed that this fact will make rational agents adopt P. Bargaining with another agent about how to jointly act in a PD will still result in people acting to defect. People care about the satisfaction of their preferences, and short of assuming that everyone cares about X (which we cannot do) the fact that the exclusion of tuistic preferences helps ensure that the chosen equilibrium is optimal does not even start to gain plausibility.

Susan Dimock argues only for the exclusion of positive tuistic preferences from bargaining situations.[62] She discerns that the satisfaction of tuistic preferences would be considered as part of the distributable surplus of any cooperative endeavor only if persons concluded that parties to any cooperative endeavor had a claim to the future benefits that in part depend upon the distribution of benefits of the first claim. This is to say that only if future benefits derived from a past transaction were possible considerations for deciding how to distribute a cooperative surplus would tuistic preferences be included for consideration. Given that any tuistic satisfactions (or frustrations) would only obtain *after* the original surplus has been divided, it ought to be considered a separate future benefit, as opposed to a benefit directly tied to the original cooperative endeavor.

If this principle of bargaining is admitted, then silver miners, for example, would be able to claim that the benefits to be divided among necessary participants to the mining of silver include not only the amount of silver received[63] but also the benefits derived from one miner crafting his share of the silver into candlesticks at some later date. The tuistic satisfaction that person A takes at person B's satisfaction occurs only after the mining of the silver, and the division of the surplus, has actually taken place, and as such is clearly a future benefit. She argues persuasively that allowing this kind of claim encourages free-ridership, and a decline in rational cooperative endeavors. And so she concludes that rational agents have reason to exclude future benefits from the domain of the present cooperative surplus to which each participant lays some claim.

Whether or not one finds Dimock's reasons compelling, what is of interest here is the form that her argument takes. She argues that rational agents, coming to an agreement over the correct understanding of what is to count as the cooperative surplus of a particular bargain, would likely not decide to include expected future benefits. This argument is most easily understood as of a contractarian type. It shows not that positive tuistic concerns should be excluded from a contractarian account, but instead that they should be excluded, due to contractarian concerns, from an account of rational bargaining. In fact, given the importance of tuistic considerations to this particular argument for what is to count as the cooperative surplus of rational cooperative bargains, Dimock's line of reasoning seems to argue for the inclusion of tuism in contractarian efforts. To not consider tuistic preferences would be to leave unanswered a question of

---

[62] Susan Dimock, "Defending Non-Tuism", *Canadian Journal of Philosophy,* 29:2, 1999, pp. 251-274. She does promise to argue for the exclusion of negative tuistic preferences at a latter date. This promise has not, to the best of my knowledge, been fulfilled as of yet.

paramount importance in rational bargaining, to wit, do we include the satisfaction of tuistic preferences in our cooperative surplus?

Having thus dismissed the important economic considerations, we have only the issue of calculability left to consider. Should we exclude tuistic concerns if they render the contractarian calculation impossible? Given our project, the answer is clearly 'yes!' Since we are here attempting to construct as motivating an account of morality as possible, anything that renders the calculation impossible must be excluded. Whether the calculation that remains possible has any merit is a question that we can answer only once the construction is complete. Do tuistic preferences render the calculation impossible? Not in general, but in specific cases some combinations of tuistic preferences render the calculation impossible. For example, any two (or more) peoples' preferences that mutually refer without end render the calculation impossible; if A's 2nd preference is that B's 1st preference is satisfied, B's 1st preference is that C's 5th preference be frustrated, and C's 5th preference is that A's 2nd preference be fulfilled, no terminus may be produced.[64] Does this justify the assumption of non-tuism? It does not. There is no reason to suppose that tuistic preferences cyclically refer in general. Does this justify excluding particular combinations of preferences? Yes it does. The same will be true of any combination of preferences that render the calculation impossible. And this leads us to consider the formal conditions that each agent's preference set must satisfy.

---

[63] Or more correctly, the satisfactions which each person would enjoy upon receiving some amount of silver.

I.i.b: Coherent Preferences

We will here take no interest in developing a correct account of the true nature of preferences, nor the correct devices through which to access them. It is beyond the scope of this project to take issue with the economic derivation of preferences by reference to choices made, or, alternatively, by way of expressions of pairwise preferences, or perhaps by way of a more direct derivation via cognitive science. We stipulate, for the reasons made clear directly above, that the sets of preferences to be made use of do not co-refer in such a way that they will infinitely cycle. This is the only inter-set requirement necessary. We will now delve into the necessary conditions that any set of preferences must meet.

We take as our starting point that we have accurately managed to capture each concerned person's preferred ordering over the set of outcomes; that we have accurately ranked each person's preference for a world in which people do not hit people with sticks, vs. one where people do not hit others with steel poles, etc.. We stipulate that this set is finite. In order to ensure that this set of preferences be applicable to the set of rules each agent will later be required to rank order, we further stipulate that this basic ranking follow the conditions set out by Luce and Raiffa in *Games and Decisions*.[65]

Agents suppose that each set of rules will result, with some degree of certainty, in certain possible states of affairs obtaining. The set of rules that includes only a prohibition on hitting people with sticks, for example will be thought by some agent to

---

[64] As has been pointed out to me by Dave DeVidi, this is importantly different from simply mutually referring preferences. If A's 2nd preference is that B's 1st preference be satisfied, and B's 1st preference is that A's 2nd preference be about agent B's 1st preference, the preferences co-refer, but a terminus results.

result in the outcome that people will not hit people with sticks with probability p, but will also be thought to bring about each other outcome (and indeed, each combination of outcomes) with some designated probability. It is imperative, therefore, that the basic preference ordering be such that it can meaningfully rank complex choices under uncertainty. The Luce and Raiffa conditions, together with a preference ranking over the set of the basic states of affairs, ensure that this ranking may be accomplished. These conditions also ensure that we can rank order these preferences over the various sets of rules not only ordinally (1^{st}, 2^{nd}, etc.), but also cardinally (on an interval scale). This becomes important when we search for a meaningful way to compare the various agents' preferences over the same set of rules.

Before discussing the formal conditions with which we measure complex choice under uncertainty over various complex states of affairs, we must to further delineate what is meant by the set of outcomes. This set contains all of the discrete outcomes that are taken as relevant results of rules regarding agent activities. We formally identify this set as $O = \{o_1, o_2, o_3, \ldots o_n\}$. Since there seems no compelling reason to define this set subjectively for each agent, this set is defined objectively, and is the same for each agent. With no obvious philosophical loss we have gained significant formal simplicity. We further stipulate that each member of O be logically independent of all others. Failing to so stipulate would allow an agent to rank a contradiction possible or a tautology as not following from some set of rules; it would further frustrate our efforts to consider the power set of these outcomes without contradiction. As these outcomes are introduced in order to gauge the effects of various systems of rules, it is appropriate to insist that these

---

[65]D. Luce & H. Raiffa, *Games and Decisions: Introduction and Critical Survey*, John Wiley & Sons, Inc., New York, 2^{nd} printing, 1958. See esp. pp. 23-29.

outcomes deal only with the patterns of behavior in which agents could engage. In particular, each outcome will be defined as a large percentage of the population refraining from a certain activity. To anticipate briefly, an agent considering the state of affairs $\{o_1\}$ is, in effect, considering the state of affairs where agents do not engage in any constraints other than o_1, and so all relevant patterns of behaviors are thereby taken into account. Lastly, and perhaps trivially, O will not include patterns of behaviors that do not effect the population of agents being considered. The patterns of behavior of the people of HD 83443 B do not affect us, and therefore do not concern us.[66]

These conditions, while fairly restrictive, are not sufficiently restrictive for O to be finite. We must further suppose that we can partition actions into sufficiently similar sets, and eliminate trivial differences of degree. The state of affairs in which most people throw a baseball at kittens at 95 MPH would be little different from the one in which most throw at 94.9882 MPH, and may both be classified as the same outcome. It is stipulated that this set of outcomes is finite. Allowing an infinite set renders the contractarian calculation impossible to conclude. While we are attempting to render the maximally motivating contractarian project we must engage in a calculation that will terminate, for the reasons made clear above.

We return now to the formal conditions regulating choice under uncertainty. Luce and Raiffa provide an accessible account of the necessary conditions for this to be ensured. To wit: 1) The preference (>) or indifference (=) ordering holds between any two outcomes, and is transitive. In other words, $o_x > o_y$, $o_y > o_x$, or $o_x = o_y$, for every pair of outcomes, and if $o_x > o_y$, $o_y > o_z$, then $o_x > o_z$. 2) Compound lotteries over outcomes

[66] HD 83443 B is a planet circling star HD 83443.

can be reduced to a simple lottery according to probability calculus.[67] 3) The value of each outcome o_n is indifferent to exactly one lottery over two outcomes o_c and o_x, when o_c is more preferred to o_n and o_x less preferred to o_n. 4) If an agent is indifferent between o_b and o_c, then in any lottery L_i containing o_b as a possible prize, o_c may be substituted to create L_x and the agent will be indifferent between L_i and L_x. To illustrate: if I like $5 as much as I like a certain lamp, then I will be not prefer engaging in a gamble that has a 50% chance of my receiving $5 to that same gamble with that lamp as a prize instead of the $5. Nor will I prefer the gamble with the lamp as a possible prize to the gamble with the $5 as a possible prize. 5) Preference and indifference among lotteries are transitive relations. 6) A lottery L_b is preferred or indifferent to L_c, given that they have only the same two possible prizes, if and only if the probability of the more preferred prize is the same or greater in L_b. These conditions are jointly sufficient for our enterprise.

I.ii: Beliefs

The conditions put on beliefs are much less complex, but are no less rigorous for that. Each agent has beliefs regarding the likelihood that each complete system of rules of interaction results in each combination of outcomes.[68] For each examined system of rules, each agent has fixed views about the probability that each member of the power set of O will obtain. This is to say that each agent has views about the probability that $\{o_1\}$ results (and nothing else), that $\{o_1, o_2\}$ results, that $\{o_{23}, o_{34}, o_{57}\}$results, through to the probability that $\{o_1 \ldots o_n\}$ results. We shall call the power set of O the set of states of

[67] Imagine a coin toss resulting, in the case of a "heads" result, in the chance of receiving a 50% chance at $1 vs. nothing and, in the case of a "tails" result, in an 80% chance or receiving a donkey vs. a horse. This conditions states that an agent would be indifferent between that lottery and a single-stage lottery with a 25% chance at receiving a dollar, a 25% chance a receiving nothing, a 40% chance of receiving a donkey, and a 10% chance of receiving a horse.

[68] As will be made clear below, a complete system of rules is any system which, for each action, permits, forbids, or obliges it.

affairs (SA) = $\{s : s \subseteq O\}$. Formally, for each rule, each agent has beliefs of the form: $P(R_j \Rightarrow s_h)$ for each $s_h \in SA$. Considering the power set has the advantage of allowing a more accurate reflection of each agent's beliefs regarding joint probability, given that it is well documented that people in general do not deal with probability well. People tend, for example, to assign the joint probability of X and Y (P[X + Y]) occurring inconsistently with their discrete evaluations of the probabilities of X occurring, and Y occurring (P[X], P[Y]). Given that we have no reason to prefer the one ranking to the other – to suppose that one is the agent's 'real ranking' - we include them all.[69]

The highly subjective content of such conditions will not sit well with the more paternalistically inclined. Many people would benefit enormously if they relied upon some set of objective facts about the world, including the various implications of each system of rules. Compelling cases can be made for the inclusion of reliable objective information, should it be available, but I do not suppose that it is so clear that such information is available.

Objective information involves facts; what one *calls* objective information usually amounts to information that one believes, and strongly believes that one has received from an authority on such matters. This information is, at least usually, not universally agreed upon by experts, and in the cases relevant to our inquiry, where this set differs from the subjective set of some agents, is not agreed upon by definition. Usually each proposed set is not even agreed upon by any reasonably defined set of 'the experts in the field.' The usual trouble with defining the set of reasonably accepted facts, together with

---

[69] This line of reasoning may be thought to also apply to the basic preference ranking over outcomes: we ought to gauge preferences over SA instead of considering simply the set of O. This ranking would render the calculation necessary here impossible, due to it, in general, failing to meet the Luce and Raiffa

each person's preference for relying on their own experience and beliefs, compels me to present, albeit reluctantly, a subjective account of beliefs.

It may be objected that this will likely result in a set of rules being relatively favored by an agent due to his mistaken impression that the outcomes he prefers are likely to occur. The facts regarding the probability of various outcomes obtaining may be quite different from his expectations about it. This is, of course, a consequence of our construction. It is true that the world would not have the values that the agent imagined above supposed, but we are here presenting the most compelling contractarian calculation, and are not unduly concerned with such a possibility. It is also possible that agents *en masse* make the same mistake, and end up implementing an impracticable set of rules. Such a possibility would raise concerns if we were here considering the stability of the rules endorsed via contractarian assumptions, as a mistake regarding the probabilities of certain results would become apparent upon implementation, and would then undermine the choice made. It is not here suggested that stability is of little consequence, but at this stage it is inappropriate to allow this concern to have any force in our construction of the contractarian enterprise.

I.iii: Rules of Interaction

We begin with a basic set of rules of interaction, RI. Each member of the set of RI is a rule that obliges, permits, or forbids agents to perform a particular action, which we write as rOa, rPa, and rFa respectively for the action a. Each of these actions is to be understood as analogous to the outcomes describe as the members of SA. They are also defined objectively, are logically independent of all others, and they concern the actions

---

conditions. Moreover, in the cases where the Luce and Raiffa conditions were met by the ranking of SA, considering that ranking would be equivalent to simply considering O.

of relevant agents in general. Further, it is a finite set into which an infinity of particular discrete actions, whose differences are trivial, are partitioned. The set of actions is formally $A = \{a_1, a_2, \ldots a_n\}$. So understood, then $RI = \{r: \exists a \in A\ (rOa \vee (rPa \vee rFa))\}$.

From RI we take as salient the sub-sets that have a rule referring to each action, and that have only one rule referring to each action. We will call each of these sets a complete system of rules of interaction, R_h. The set of complete systems of rules is SR. That is, $R_h \in SR$ if and only if $\forall a \in A, \exists r \in R_h \{[rOa \vee (rPa \vee rFa)] \wedge \neg\exists r' \in R_h [(rOa \wedge (r'Fa \vee r'Pa)) \vee (rPa \wedge (r'Fa \vee r'Oa)) \vee (rFa \wedge (r'Oa \vee r'Pa))]\}$. It may strike one as odd that we require that each $R \in SR$ not have any action that is both permitted and obligatory, given that if some action is obligatory, deontic systems of logic require that that action is also permitted. We invoke this restriction so that what would, practically speaking, amount to equivalent systems need not be considered twice by various agents. With SR in hand, we may now proceed to outline the contractarian calculation. It proceeds in two stages: first at the level of individual choice and then at the level of social choice.

II: The Contractarian Calculation

The following account of the process of individual choice and then collective choice over the proposed alternative complete systems of rules proceeds in a fairly non-specific fashion. We propose to discuss in the most general terms what it is that the contractarian calculation must accomplish for the ultimately compelling result here desired to obtain. A more detailed analysis will follow when we examine the calculability of the contractarian calculation.

II.i: Individual Choice

At the level of individual choice, we must first assign each $R_x \in SR$ a cardinal utility value. For each relevant agent we have at our disposal beliefs regarding the probability of all possible combinations of outcomes given each possible combination of rules. We also have at our disposal a ranking of the various outcomes. This ranking has been constructed in such a way that it meets conditions sufficient for various complex lotteries to be meaningfully compared and evaluated. Each agent's belief set about the outcome of any particular member of SR can be presented as a complex lottery over outcomes. For each $R_x \in SR$, the agent's beliefs regarding the likelihood of each combination of the various outcomes are retrieved from her set of beliefs, which, it will be recalled, are of the form $P(R_x \Rightarrow s_h)$. The set of beliefs for each R_x, $\{P(R_x \Rightarrow s_h), \ldots, P(R_z \Rightarrow s_m)\}$, in effect identifies a lottery which, if chosen, results in a chance of P_a that $\{o_1\}$ and a chance of P_b that $\{o_1, o_2\}$, etc. Given each agent's preferences over outcomes, then, once this process of belief retrieval is concluded for each $R_x \in SR$, we have a set of complex lotteries which is to be ranked via each agent's original preference ranking over outcomes. These steps are to be repeated for each relevant agent.

II.ii: Social Choice

We are now prepared to make the leap from individual choice to social choice; to infer what the chosen set of rules for interaction must be, given individual preference rankings. In making the preparations for such a construction, one must avoid constructing a model that falls victim to Arrow's theorem: that there is no voting model

that will satisfy connectivity[70] and transitivity, while allowing 1) that individuals are free to order sets in any way that they choose, 2) that social ordering satisfy Paretian value judgments (if even one person prefers X over Y and none prefer Y over X, then X is preferred over Y in the social ranking), 3) that the social ordering ranks each pair (X,Y) depends solely on X and Y, and not on any other possible outcome Z, 4) that a social ordering shall not be imposed on individuals, and 5) that the social ordering does not depend only upon the ranking of any one person regardless of the others.[71]

We will consider two models of social choice available to those wishing to avoid the pitfalls of Arrow's theorem. The first is Gauthier's proposal that individuals engage in bargaining with each other over a joint strategy that will consider the optimal sets of rules, and ultimately decide which is to be endorsed by all. We will find this first alternative inadequate. We will embrace the second proposal. Finding Gauthier's heterodox account lacking, we will follow the orthodoxy, constructing an account which makes use of the relative strengths of people's preferences – a route available to us due to our insistence that each subjective scale be measured not only ordinally, but cardinally as well. This further requires that we find some criteria with which to determine the relative weight of each person's preference set. We suggest that the project requires counting each person equally.

The rejection of Gauthier's solution depends on no particular feature of his theoretical efforts; the reason to reject his proposal is sufficient to reject any such attempt. Any suggestion that bargaining may be used entails that one already has a

---

[70] Connectivity is the completeness of the social ranking of preference or indifference over each pair of choices.

[71] In this description we follow D. M. Winch's *Analytical Welfare Economics*, Penguin Books Ltd., (Harmondsworth; England) 1971, pp. 176-180.

starting point from which to begin bargaining. People engaged in bargaining have certain attributes or possessions that they may bring to the table, certain offers that they can make that have particular values to the other individuals involved in the proposed bargain. It is only in such a way that others may come to conclude that it is rational to engage in such-and-such a bargain, given what it is that A has offered to contribute, and given what A demands as compensation for that contribution. In the most familiar cases of bargaining, that which A has a right to offer to B and C is already well understood, as the expected relative value of that offer to B and to C in the future (post bargain). All this can also be said of formal theories of rational bargaining: a baseline must be presupposed.

In this particular instance, however, what may be offered, and the respective expected values of that offer, depend in large part upon the rules which are eventually chosen. Should computers be banned, it will likely turn out that we each value Bill Gates less than we otherwise would; his long term prospects seem less good. We would devalue him even further if it is the case that people do not retain title to what they have previously acquired or produced. The support of the physically strong while walking through dark alleys will be less significant if we are all required to carry fully automatic weaponry. In short, without an understanding of the relative worth of each person's contributions in the future, or even an understanding of what that person's contributions are likely to be, no bargaining can take place. Without such a starting point from which

to make offers, the offers could not be rationally accepted. Bargaining about the joint strategy that is to choose between optimal points, then, is worthless in this instance.[72]

Gauthier's proposed baseline rests on an extension of his interpretation of the Lockean proviso: each may make use of that which is exclusively theirs, and anything obtained with the same, provided that none were taken advantage of during such procurement. While we will more fully explore his positions in chapter 5, it is worth noting that his particular defense falls prey to the above objection. If his Lockean proviso is to determine each person's worth in the bargain, then it must be able to determine not only what possessions people are going to retain, but also what their powers and possessions are worth to others. This cannot be determined without some expectation that people will retain title to what they have acquired pre-bargain, and that there is some basis to expect that the strategic importance of each possession can be known in advance of the bargain. There is no reason to expect the former, and no reason to assume the latter.

We return to the more orthodox account mentioned above: making use of the relative strength of people's preferences over the various alternatives. As is well recognized in economics literature, Arrow's theorem depends in part on each agents' ranking over various alternatives being ordinal.[73] Deriving social choice directly from individual choice in this fashion demands a non-arbitrary scale with which to compare different agent's strengths of preference over alternatives. A cardinal measure of the strength of a person's preference for outcome O_n is determined relative to the set of an

---

[72] An introduction of any 'state of nature' fixed point that stipulated what each can expect future contributions to be worth would clearly overcome such an objection, but any such introduction would also clearly be an example of question begging.

[73] See, for example, Winch, 1973 pp. 175-189.

agent's preferences over all alternatives. The measurement tells us nothing about how it compares in strength to another person's preference for O_n. In order to produce an index on which to compare interpersonal utility, we must find a principled method by which we can weight the worth of each person's set of preferences.

Given the nature of this enterprise, one obvious suggestion to make at this point is that people's preferences over the set of possible systems constraining interaction be weighted in relation to each person's market value – the value that others place upon their participation in the system being agreed upon. A person's preferences count to the degree to which others instrumentally value the expected contributions of that person. Unfortunately, this proposal is untenable for reasons given above. Given that a person's contribution depends upon what is considered her property, and what it is expected that she would do with that property, any meaningful expectations about her contributions are here impossible to predict. An expectation about a person's contributions must depend in large part on which rules are chosen to govern interaction. Given our current position, then, such dependence negates the applicability of such considerations as expected contribution to our enterprise.

It is worth reconsidering what, exactly, our enterprise is, in order see what consideration could both be consistent with it, and recommend an index that will allow meaningful interpersonal utility comparisons. We are here undertaking to produce a schematic that will outline the procedure that must be followed in order for the most rationally compelling contractarian justification of a system of constraint to be produced. We have not, as of yet, discussed in much detail what it would be for a contractarian justification be 'most compelling'. This phrase invites misunderstanding. We could

mean 'the most compelling' to agent A_j. Such a justification presumably would amount to an identification of a set of rules found most favorable by A_j. Or perhaps such a justification would amount to an identification of a set of rules most favorable to A_j which agents A_1 to A_n (the rest of the agents being considered) would adhere to if it were to be implemented. It is assumed obvious that this is not what we here mean to present. The most compelling set of rules is to be understood as a set of rules which equally compels each agent. Given that we have no interest in convincing agent A_j to any greater degree than agent A_c, or any other agent, we give no more weight to A_j's preferences than any other agent. In attempting to compel each agent, one must appeal to each agent's interests equally, and so we suggest that each person's cardinal preference ordering be scaled between 0 and 1, so that an interpersonal comparison is possible.[74]

Given this conclusion, then, the final step of the contractarian calculation is immediately evident. Having ensured a common measure, we simply add the weights of each agent's preferences for each suggested rule governing interaction, and that state of affairs which is thus accorded the greatest number is said to be chosen.[75] Attempting to convince a population that a given set of rules is the one which all should adopt, we could offer into evidence that people felt most strongly about this particular set of rules. Any attempt to implement any other set of rules, then, would lead to at least as much resistance as to this set, and likely quite a bit more.

It is worthwhile to note that amalgamative contractarianism and bargaining contractarianism (should this latter be practical) would likely identify the same set of

---

[74] Although we present this assumption here, we formalize it by ranking each agent's basic preferences over outcomes between 0 and 1, as suggested by Luce and Raiffa.

[75] There is nothing that suggests that ties are impossible. We suppose that a randomizing device would be used to break ties.

rules under some general conditions. If there is a set of rules that would be most strongly preferred by each agent to any other set of rules, then this set of rules would obviously be identified both by the amalgamative and bargaining model. While bargaining, each agent with all others, each agent would be a proponent for the same set of rules governing interaction, and the bargain would then obviously result in that set being chosen. In the amalgamative model, if each person preferred one particular set of rules (R1), then that set of rules would receive the highest numerical value in each agent's individual ranking of all of the members of SR. Summing all agents' preferences for each set of rules, the sum for R1 would have to be the greater than any other sum. Insofar as there exists a set of rules found to be unanimously preferred over all others, the results of these two versions of contract theory would not differ. Theorists who are convinced of the existence of such a set, then, have no reason to fear that amalgamative contractarianism would fail to identify that set of rules as authoritative. Those theorists who are not so convinced will take issue with the amalgamative contractarian's assertion that a system of rules not most preferred by some agent could meaningfully be said to be the 'rational' set of rules for that agent to adopt. To these theorists we have to reiterate that, given that the bargaining model of contractarianism is impossible to implement, the amalgamative option is all that remains to be attached to this term.

III: Calculability

In this penultimate section of the chapter, we will examine what chance the contractarian has of rendering any timely result from this calculation. Given that one ultimately wishes to compel the agents for which this calculation has been run, and given the average life expectancy of people today, we could suggest that 'timely' ought to be

understood intuitively as 'less than 76 years'. This understanding entirely ignores complications such as the birth and death rate, which would likely result in a different set of agents having the results presented to them than was included in the calculation. Similarly, this ignores completely that people's beliefs and desires quite commonly change over the course of their lives – perhaps even day by day. Making an effort to include these complications in the project renders it unlikely in the extreme that a result could be obtained in a timely fashion. It is entirely unclear what a 'timely result' would then amount to; and we suspect that it would demand that a result obtain almost instantly.

Thankfully, we do not have the task of delineating what, exactly, a timely result would amount to. Appealing to computation theory allows us to avoid getting bogged down in such a complicated matter. Computation theory understands all deterministically solvable problems, which are solvable in polynomial time, tractable. Those problems not solvable in polynomial time are considered intractable. Any intractable calculation may be considered to not result in a timely solution. Intuitively, a problem that increases in complexity on a greater than polynomial scale is going to be far too time-consuming in all but trivial cases. A problem that requires that 2^n steps – such as the construction of a truth table - will require 68,719,476,736 steps for a sentence of 36 characters. Any problem that is solvable in polynomial time, that is, a problem which requires steps of n^c for some constant c, is not thought to become too time-consuming for greater than trivial cases. This is not to say that any tractable problem will ultimately turn out to be calculable in practice. Given a very large constant value, some calculations will be tractable, and nevertheless not be possible right now. A calculation where n = 4 and c = 300 requires 4.14951556888099295851240786369 1e+180 steps. Given that we do not

intend to explore exactly how many rules of interaction are to be considered, nor how many agents we will consider, nor how large the set of beliefs of each agent actually is, etc., we can only here discuss computability at the level of tractability. We leave unexplored more fine-grained evaluations.[76] We will ultimately show that the contractarian calculation is tractable, but we must first proceed to show that the calculation is possible.

Both problems deterministically solvable in polynomial time (P) and those thought not deterministically solvable in polynomial time, but for which a proposed solution can be checked in polynomial time (NP), are sub-set of the class of calculable equations.[77] Without showing that it is possible to deterministically solve the contractarian calculation given the inputs suggested above, any further examination is pointless – if it is not calculable, it is not P or NP, and so definitely not P. We will show that each calculation necessary to perform the above-described calculation is a member of the class of *primitive recursive* functions. Primitive recursive functions are calculable; in fact, the set of primitive recursive functions is the most basic set of calculable functions. Any finite serial application of primitive recursive functions is also a primitive recursive function, and therefore calculable. So in showing that the contractarian function is composed of primitive recursive functions, we will have also shown that it is calculable.[78]

---

[76] I would like to thank Dr. J. Shallit for pointing out the importance of this fact to my proposal.

[77] We will be throughout assuming the plausible conjecture that $P \neq NP$.

[78] The following account of primitive recursive functions relies quite heavily on H. R. Lewis & C. H. Papadimitriou's proofs in *Elements of the Theory of Computation*, Prentice-Hall Inc., (Englewood Cliffs; New Jersey) 1981, esp. pp. 232-248.

A primitive recursive function is one that can be generated from the following basic set of initial functions by way of repeated applications of *composition* and *primitive recursion*:[79]

1. The *0-place function.* ζ represents the function from N^0 to N such that $\zeta(\,) = 0$.[80]
2. A *projection function.* A projection function, π, is one that, for each $1 \leq j \leq k$, $\pi_{jk}(n_1, \ldots, n_k) = n_j$.
3. The *successor function.* The successor function, δ, is the function from N to N such that $\delta(n) = n + 1$ for each $n \in N$.

A function, f, is *composed* of other functions g and $h_1,\ldots,h_j$ if $f(n_1, \ldots n_k) = g(h_1(n_1, \ldots n_k), \ldots h_j(n_1, \ldots n_k))$. When *g* is a *k*-place function, and *h* is a (*k*+2)-place function, and *f* is a (*k* + 1) place function such that for every $(n_1, \ldots n_k) \in N^k$, $f((n_1, \ldots n_k), 0) = g(n_1, \ldots n_k)$, while for every $m \in N$, $f((n_1, \ldots n_k), m + 1) = h((n_1, \ldots n_k), m, f((n_1, \ldots n_k)))$, then *f* is said to be derived from *g* and *h* by *primitive recursion.* Intuitively, the three functions described above are calculable if any functions are, and any combination of these functions via composition or primitive recursion, being built up of computationally trivial building blocks, will be calculable as well. We will not dwell on these details or upon proofs that certain equations are primitive recursive; we simply note that summing and product functions are primitive recursive functions.[81]

A predicate, or relation on a subset of natural numbers, is primitive recursive if and only if its characteristic function is primitive recursive. A characteristic function of a k-place predicate is that k-place function *f*, that maps N^k to {0,1} such that $f(n_1, \ldots n_k) = 1$

---

[79] This trivially implies that the set of initial functions are also primitive recursive.

[80] N will hereafter represent the set of natural numbers, N^k represents a Cartesian product of the natural numbers k ordered elements long, and N^0 is the empty set.

if $(n_1, \ldots n_k)$ has the relation, and $f(n_1, \ldots n_k) = 0$ if it does not. Foregoing the proof for the sake of a linear presentation, we merely state that the equality relation, less than relation, and the less than or equal to relation are all primitive recursive.

At the level of individual choice, we must present a cardinal utility ranking over each $R_x \in SR$. The first function we make use of calculates the utility of each $s_h \in SA$. It will be recalled that each s_h is composed of a set of outcomes over which we have a cardinal utility ranking. The utility of each s_h is equal to the sum of the utilities of the outcomes that comprise it. Formally, $f(s_h) = u(o_1) + \ldots u(o_n)$, for $\forall o_m \in s_h$. We then proceed to assign a utility measure to each $R_x \in SR$. For each rule, the expectation that each s_h obtains ($P(R_x \Rightarrow s_h)$) is multiplied by the utility calculated by $f(s_h)$. This provides us with each s_h's expected utility, given the R_x being considered. These results being summed provide us with each individual's ranking of R_x. This is repeated for each $R_x \in SR$ to provide us with the individual's ranking of each of the proposed systems of rules. Repeated for each agent, these calculations provide us with all considered agents' rankings of the proposed systems of rules. As these calculations are all known to be primitive recursive, the contractarian project is now known to be primitive recursive at the level of individual choice.

At the level of social choice the ranking of each R_x given by each individual is summed together with every other agent's ranking of the same. This set is then ordered according to a "greater than or equal to" relation. The rules with the highest overall sum are appropriately chosen. As this stage of the contractarian calculation is also composed

---

[81] The enthusiast is referred to Lewis & Papadimitriou's proofs in *Elements of the Theory of Computation*, 1981, esp. pp. 234-239.

of functions known to be primitive recursive, the entire calculation is thus now known to be calculable.

Moreover, this calculation is also clearly calculable in polynomial time. The addition of each agent merely ensures that the calculations at the level of individual choice must be run one more time, and that at the level of social choice, the ranking of each of the rules being considered has one more number to be added to the sum of the others. Following the conventional wisdom of computer science, then, we deem the contractarian calculation calculable in a timely fashion, given the appropriate inputs.

IV: An Objection Considered

A critic may inquire of our investigation whether we have assumed too much at the outset. Gathering our assumed inputs is at least incredibly complex, if not impossible to perform. Gaining accurate access to a person's basic desires and beliefs is, admittedly, not unproblematic. And this is to say nothing about how to manipulate a person's preferences over outcomes in such a way as to have them conform to the Luce and Raiffa conditions; we will examine this presently. But while the possibility of gaining this access is a complex issue, it is not our issue. The philosopher cannot be expected to compete with an economist, or psychologist, or cognitive scientist, in their respective areas of expertise. Knowing of no impossibility proof against the successful completion of each of these possible methods of gaining the necessary information, then, we content ourselves with the optimistic assumption that this access is possible.

However optimistic this assumption, it does nothing to cheer the contractarian who wishes to provide a contractarian justification *now*. Forbearing from recommending

a termination of all contractarian activities too hastily, we turn first to examine three possible routes down which the contractarian might turn instead of coming to a full stop. We first turn to Herbert A. Simon's satisficing agents in an effort to see what may be said for making use of agents with a simpler cognitive process than straightforward maximization. Secondly, we examine the work of contractarianism's most famous proponent, Thomas Hobbes. We there focus on what may be said of the practice of only identifying a few preferences that almost all people have, and have quite strongly. Finally we turn to the works of David Gauthier to see what benefit there is in considering agents entirely devoid of particular preferences, and identifying which rules of interaction favor economic interests in general. It is through the conclusions drawn from the examination of these particular turns that we hope to console the contractarian who wishes to present a contemporary contractarian justification, but who is aware that the requisite input is, thus far, lacking.

CHAPTER 3
~ Satisfactory Contractarianism? ~

Contractarians who have proceeded this far, and who nevertheless wish to attempt to produce a contractarian justification in spite of the impossibility of currently producing an exhaustively informed effort, must examine the suggestion presented thus far to find what is lacking. We can separate the previous proposal into two basic units: the agents and the calculation. It has been shown that the calculation is not problematic. The problem lies with the constructed agents. These agents were constructed in such a way as to be exhaustively informed maximizing agents. Finding such maximizers inadequate for the above-proposed purposes, we proceed to examine the most commonly proposed alternative: satisficing agents. Satisficing agents are widely regarded as the obvious alternative for instrumentalists about practical rationality who, for whatever reason, find the traditional maximizing account lacking.[82]

I: Satisficing Agents

Satisficing agents are traditionally proposed as alternatives to maximizing agents in economic analyses – most famously by H.A. Simon, who first proposed that making use of satisficing agents would produce more accurate economic analysis of administrative behaviors than would reliance upon optimizing agents.[83] These satisficing agents are typically envisioned as failing to exhaustively consider all known (or knowable) alternatives in a given choice situation in an effort to bring about the optimal

---

[82] See, for example, Michael Byron's "Satisficing and Optimality", *Ethics*, 109 (October 1998), pp. 67-93, esp. ff. 67.

[83] Most famously in H. A. Simon, *Models of Man*, John Wiley & Sons, Inc., (New York) 1957, esp. pp. 241-260.

state of affairs. Satisficers instead search for an alternative that is seen to be 'good enough', and act accordingly.

This brief description does not accurately capture the original intuitions of Simon, as he can be taken as originally supposing that satisficing agents do not even engage in deliberative behavior when acting in market transactions, given their total lack of access to other agents' subjective belief states.[84] Nor does it capture any general tendency in the field of heterodox rational choice to produce satisficing agents of a particular type. There is a tremendous amount of divergence regarding the proper construction of a satisficing agent. There are proposed satisficing agents who knowingly choose a lesser over a greater good (anti-maximizer); who satisfice as a meta-strategy in an effort to efficiently conclude any given deliberation; who satisfice simpliciter (not as a meta-strategy); who have access to all possible solutions, but only choose to examine 'promising' alternatives; who do not gather all possible solutions, but exhaustively examine those solutions known; etc. Perhaps the only defining feature shared by each of the agents is what they do not do: they do not always act in such a way as to exhaustively enumerate and compare all possible solutions to a given problem in an effort to identify the optimal solution.

The plethora of available alternatives to maximizing agents may be in part explained by the variety of disciplines from which such agents are derived, and the respective goals of each such discipline. Economics has taken its cue from Simon, and in the main has attempted to construct satisficing agents for the purposes of more accurately

[84] H.A. Simon, 1957, pp. 259-260.

modeling economic behaviors.[85] Such activity is best understood as a reaction to neoclassical economic analysis, which assumes perfectly rational agents who have at their disposal perfect information regarding their environment and interests, and reason instantaneously. Computer science, on the other hand, makes use of satisficing agents in an effort to overcome obvious difficulties with what are called calculatively rational algorithms. These calculatively rational algorithms can produce the correct answer to a given problem, but given calculation costs and problem complexity, may produce the answer far too late to be of any use. As these answers are (by definition) of no use, other algorithms must be used: algorithms that produce an acceptable answer, and do so within the necessary time constraints. These algorithms are obvious analogues to satisficers. Such efforts are clearly prescriptive in nature; each proposed algorithm is recommended as a better way to solve a given type of problem. Philosophy has concerned itself with both types of projects; prescriptive approaches to rational choice scenarios of fairly complex natures are of particular interest to us here, and there is a vast literature examining the proper understanding of practical rationality.

We will be primarily concerned with spelling out the various ways in which these satisficing agents may be envisioned as carrying out their assigned procedures and what reliance on this type of agent might achieve for contractarians.[86] It will ultimately be concluded that the construction of a satisficing agent is no help to the contractarian. This result is important both because it corrects the current perception that satisficing agents *are* of some use to the contractarian (as a project involving instrumental rationality), and

---

[85] As tempting as it may be to simplify Simon's interests to solely descriptive projects, it is obvious that he also has normative ambitions. See, for example, his "Theories of Bounded Rationality", *Models of Bounded Rationality: Behavioral Economics and Business Organization*, Vol. II, MIT Press, (Cambridge:Massachusetts) 1982, pp. 408-423.

because it will further clarify what it is that the contractarian must attempt to change in his current approach.

II: Slote's Satisficers

In an effort to clear the field as much as possible, we will first examine and reject reliance upon a fairly radical type of satisficing agent introduced by Michael Slote in *Beyond Optimizing*.[87] Slote proposes a satisficing agent that can deliberately reject that which is known to make that agent better off in favor of that which is good enough; such an agent may be thought to be an anti-maximizer. In fact, Slote wishes to propose that this kind of agent more accurately models our common sense understanding of what it is for a person to be rational, and proposes that satisficing rational choice is a viable alternative to optimizing rational choice when attempting to describe deliberative behavior. Insofar as Slote suggests that actual people use the term 'rational' to describe more than simply deliberative behavior, he will of course get no argument from any even minimally observant person. People use 'rational' to indicate a person's lack of behaviors attributable to neuroses ("Well, he's more rational now than he was when he went *into* the sanitarium") and lack of emotional influences ("He's much more rational now that the funeral has taken place") and that a person has certain cognitive abilities ("My teenage daughter is just *barely* rational – she just can't seem to understand that her actions have consequences") and that a person has reasonable beliefs about the world ("My brother's rational retirement plan does not involve winning the lottery ") and for many other loosely related purposes.

---

[86] We will hereafter be ignoring the differences inherent in projects concerned with normative vs. descriptive objectives.

[87] Michael Slote, *Beyond Optimizing*, Harvard University Press, 1989.

Insofar as Slote suggests that people are not properly understood as solely rational calculators, he will also get no arguments. At most, people are understood as entirely instrumentally rational agents as a simplifying assumption when analyzing market behavior – in order to make sense of deliberative choice and market forces, one first assumes that all choice in a market is deliberative. Everyone acknowledges that people are a mix of affect, automatic behaviors and deliberative behaviors; the dispute centers on the ratio of each.

Slote, however, maintains a more radical line:

> [B]y exploring our intuitions about rational choice I hope to show that choosing what is best for oneself may well be neither a necessary nor a sufficient condition of acting rationally, even in situations where only the agent's good is at stake.[88]

> I think it can also be shown that it sometimes makes sense deliberately to reject what is *better* for oneself in favor of what is *good and sufficient* for one's purposes[89]

He attempts to make his case with a series of examples that are designed to suggest that: 1) we commonly understand as rational behavior a class of actions where the better is rejected in favor of the less good because the less good is perfectly satisfying, and 2) we commonly understand maximizing behavior as irrational. These examples take on three basic forms: ones in which we reject a known better for a known good enough, ones in which not being satisfied with a good enough option seems peculiar, and ones in which choosing a less than efficient means to a given end seems rationally permissible. We will examine each, find them lacking, and conclude by suggesting that his project is analytically incoherent.

---

[88] *Beyond Optimizing*, p. 1.
[89] *Beyond Optimizing*, p. 2.

Slote enjoins us to imagine a man deliberately dressing before going to work. While engaged in the project, he also glances outside to gauge the weather a few more times than necessary, gets sidetracked by a messy pile of newspaper and begins to straighten up before changing his mind and returning the paper to its original position. This is not an optimally efficient way to get dressed, Slote rightly suggests, and nevertheless we would not label such behavior irrational. And so, he concludes, our understanding of rational behavior cannot merely involve attempts to be maximally efficient. It must be rationally permissible to choose a less than optimally efficient means to our ends. And so we have some evidence that deliberative rational behavior must involve more than just considerations of maximization.

In order for this conclusion to follow from this example it must be assumed that tidying his room and checking the weather were not deliberative goals; otherwise Slote would simply be describing a case where one goal is pursued simultaneously with other goals. Once this is understood, it is clear that these alternative activities must be ascribed to affected behavior. The unconscious occasionally affects behavior during the course of a deliberative endeavor, and we do not commonly ascribe irrationality to the agent so affected. This is all well and good, but it surely does not support the conclusion that deliberative rational behavior involves more than maximizing behavior. Instead this suggests that we commonly use the term 'rational' to denote a lack of undue interruption of purposeful behavior by affected behavior. By way of illustration, let us imagine at what point we would describe Slote's agent as irrational. Would it be when he glances out the window to check the weather 5 times? When he checks the weather 500 times? When he checks only 5 times, but stares unthinkingly out the window for 10 minutes at a

time? At some point, the extreme nature of the unconscious behavior will prompt us to conclude that such an individual is irrational. We would not mean that he is too inefficient while pursuing his goal of getting dressed. Instead, we would be suggesting that he is crazy.

Slote also describes a scenario in which a person with a large group of intimate, personal friends states that he wishes to engage in some (unspecified) activity so that he may make more intimate, personal friends. Upon being told that he already has several friends who fit the bill, he responds by suggesting that since more is better, more friends will make him even better off. Slote, again rightly, suggests that we would think that this response would strike the average person as "specious and bizarre."[90] He also supposes that our reaction to this response depends on our finding it peculiar that the person isn't satisfied by the friends that he has – that he isn't satisfied by what is good enough.

I suggest instead that it is at least as likely that the oddity of the response depends on an agent's not recognizing that it is impossible to maintain large numbers of intimate friends, or that more intimate friends is not necessarily better. In short, I think that this response is considered irrational because the agent is supposed to lack certain fairly widespread and firmly held beliefs that each of us has acquired. His response is supposed irrational due to his lack of reasonable beliefs about the way that the world works. This example certainly does not yet suggest any compelling evidence that deliberative rationality, as commonly understood, can involve choosing a good and sufficient option over one that is better.

And so we turn to the third type of example through which Slote attempts to motivate us to accept his proposal. Imagine that you have just had a satisfying lunch, and

further suppose that there is a candy bar on your desk. You would enjoy a candy bar, but as you are perfectly satisfied given the lunch that you have just eaten you forgo the tasty treat. You do not forgo the treat because you are on a diet, or to not ruin dinner (in short, not because of any other conflicting desire) but because you are perfectly satisfied. You "turn down a sure enjoyment, because you are perfectly satisfied as you are."[91] And so Slote's strongest case for his alternative account of deliberative rationality goes. If we can make sense of this account, then it makes sense to talk about deliberately choosing the good enough to that which is better; if not, then Slote's last type of evidence for his position has been found lacking.

Slote himself seems at times to make sense of this example by appealing to habitual responses. He states that "[we] occasionally reject afternoon snacks... because... we have a habit of moderation."[92] If his thesis was that we sometimes act for reasons other than that we have made a deliberative choice, then this example would indeed support his case. However, if this were his thesis, he would not need to support it. As mentioned above, no one disputes this. People are commonly understood often to act unconsciously; and habitual behavior falls into that category.

What Slote must produce is an example that can coherently incorporate the idea that one is perfectly satisfied with a certain state of affairs, and so rejects what is better. Herein lies the problem. If one is perfectly satisfied, this means that all of ones desires are fulfilled – no more utility satisfaction is possible. If one is to be better off than one was previously, this means that a larger amount of preference satisfaction occurs than had occurred previously. But it is impossible to both be completely satisfied and to then be

---

[90] *Beyond Optimizing*, p. 77.
[91] *Beyond Optimizing*, p. 10.

made better off. Slote's example, then, is incoherent, as it relies both on one being perfectly satisfied and it being possible for one to be made better off. This incoherence permeates this entire proposed enterprise. The same can be said for any effort to define deliberative rationality in terms other than maximizing behavior. Deliberative behavior just *is* means/ends reasoning. Any effort to suggest that means/ends reasoning is something other than reasoning about the appropriate means to a given set of ends is bound to fall into absurdity. Insofar as one is engaged in an effort to convince people of the rationality of a particular set of rules governing interaction, one must make use of an agent who *at least* chooses the best of any examined alternatives. As the contractarian is engaged in an instrumental enterprise, no use can be made of Slote's type of anti-maximizing agents.

III: Instrumental Satisficers

Clearing the field of anti-maximizing satisficers still leaves quite a few different types of satisficing agents in view. They all share the virtue of being consistent with our proposal in virtue of the fact that they will never recommend a good known to be of lesser value over one that is known to be of greater value.[93] As discussed above, these satisficers come from a variety of disciplines, each with their own particular vocabulary. Computer science, economics, and philosophy all have very different enterprises in mind when they investigate the world, and their word choice reflects this fact. We will here proceed to examine some of the main types of proposed satisficers without going into too

---

[92] *Beyond Optimizing*, p. 13.

[93] It is worth noting at this point that by rejecting Slote's satisficing agents, we have not rejected algorithms that randomize over a set of possible answers, some of which are known to be superior to others. In the event that one of the less appropriate solutions was identified through this randomizing procedure, it could not be correctly suggested that the algorithm compared the identified solution to the members of the set. Any such suggestion would depend on an equivocation on the term 'known' for its plausibility.

much detail about the various titles ascribed to them, nor about irrelevant differences in their suggested construction. This allows us to avoid complex examinations of, say, David Schimitz's discussion of global vs. local satisficing and its relation to algorithms designed to operate efficiently given unknown time constraints.[94] Such a detailed and lengthy discussion would only obscure the presentation of fairly general reasons why each class of proposed agents is unacceptable for our purposes. That being said, we shall proceed to examine each of these agents in turn.

III.i: Meta-Rational vs. Simple Satisficers

The instrumental satisficing agents that have been proposed fall into two exhaustive classes: meta-rational satisficers and simple satisficers. Simple satisficers are satisficers simpliciter. These agents are constructed in such a way that they fail to exhaustively consider all alternatives. There is no higher level reasoning behind the satisficing approach to any given choice scenario – the agent is only capable of satisficing behavior. Meta-rational satisficers are agents who satisfice in particular situations because they have determined that this is the rational response.[95] Meta-rational satisficers are of two distinct types: those that have decided in all problem scenarios to use one satisficing strategy (rigid satisficers), and those that decide which satisficing strategy to employ on a case-by-case basis (flexible satisficers). Concerns regarding the costly nature of determining which satisficing strategy to employ in any given choice situation straight-away give rise to more complicated agents, agents who, at the meta-

---

[94] See chapter 2 of David Schmidtz's *Rational Choice and Moral Agency*, Princeton University Press, 1995 for the former, and Suart Russell & Devika Subramanian, "Provably Bounded-Optimal Agents", *Journal of Artificial Intelligence Research*, 2 (1995) pp. 575-609 for the latter.

[95] This leaves open the possibility that such agents may, some of the time, conclude that it is cost-effective to act as maximizing agents.

level, satisfice when deciding which lower-level satisficing strategy to employ. This can be seen to immediately give rise to an infinity of satisficers of various levels of complexity, as these agents may satisfice at the meta-level simplistically, or this satisficing behavior might be the result of further meta-deliberations regarding the rationality of their meta-satisficing strategy.

Thankfully, we need not examine the complexity of infinitely many satisficing agents. We can satisfy ourselves with an examination of the three types just mentioned above: simple, rigid and flexible satisficers. For our purposes, there is no discernible difference between rigid and flexible satisficers. Given that we are here concerned solely with the contractarian calculation, agents need only consider this one choice scenario. As these flexible satisficers need only consider one choice scenario, they may be reduced to their rigid counterparts – whatever the reason for adopting a given satisficing strategy, what is important for our purposes is the chosen strategy in this particular instance.

This line of reasoning clearly also allows us to identify any meta-rational agent with its simple counterpart: we are interested in which satisficing strategy, if any, will allow us to satisfactorily conclude our hitherto interminable endeavor. To put this point another way, *we* are engaging in the meta-deliberations about which satisficing strategy is appropriate. Our interest is thus properly focused upon simple satisficers.

III.ii: Static vs. Responsive Satisficers

Static satisficers are ones that, in each problem scenario, inevitably invoke the same satisficing strategy over and over again. Responsive satisficers, in contrast, given serial exposure to a given choice situation, may eventually evolve better and better satisficing responses. These agents may proceed, for example, by randomizing over

various strategies for a given length of time, and then endorse the one found to result in the average best answer. They could also proceed by mutation, sticking with a particular type of strategy, until a modification of the original algorithm is found to produce better results (changing the number of options considered out of a given field, for example) and thereupon adopting that latter strategy. Any number of responsive alternatives can be envisioned, and all are inappropriate for our purposes.

Any learning satisficer is deemed inappropriate because any contractarian result obtained by recourse to learning satisficers will likely be regarded as inferior to the result of running the calculation one more time. If approached with any timely result, a person could quite reasonably ask that the calculation be run again, as their 'satisficing' counterpart will not produce a less good result (for the person) and may indeed provide a better result. Infinite iteration is not an acceptable result; the contractarian calculation must produce an authoritative result. Reliance on this type of satisficing agent is therefore not advisable.

Closer inspection of this line of argument coupled with the conclusions reached when investigating the meta-rational satisficers are together sufficient to generate the conclusion that we ought to attend exclusively to simple satisficers. Insisting that satisficers have only one application of the contractarian calculation in which to act ensures that the agent can make use of only one satisficing strategy, and we do not care about how they come to make use of that strategy in this choice scenario.

III.iii: Simple Satisficers

Simple satisficers make use of one of two possible 'stopping rules', each an obvious extension of the pressures which here concern us: time and value.[96] A satisficer may either: 1) examine various alternative actions in a given field and pick the best option evaluated by a given time, or 2) examine one possible alternative after another until one has been found to have results which are at least as valuable as a pre-determined measure. The former may be characterized as picking the best of the available alternatives given time constraints, and the latter as choosing a satisfactory alternative. As has been so ably observed by David Schmidtz, however, "A satisfactory [option] may or may not be optimal. Likewise ... the best available ... may or may not be satisfactory."[97] This latter consideration brings us to discount from consideration those satisficing agents that rely on choosing the best of available alternatives. There is no guarantee that the people that we wish to convince of the rationality of the contractarian calculation are going to find the alternative proposed satisfactory. For each agent, the rules considered by the relevant satisficer could have been the worst rules possible among the members of SR. None of these rules would be acceptable to that person. People, we submit, could rationally endorse no calculation that allows for this possibility. And so we turn to satisficing agents that search for acceptable alternatives, and terminate once one is found. Those agents have set "aspiration levels" that a solution must meet to be deemed an acceptable termination point.

The issue before us, then, is obviously how to properly set the aspiration level of each agent. Aspiration levels are going to be set either objectively or subjectively. That

---

[96] We will hereafter use the term satisficer to denote simple satisficers.
[97] David Schmidtz, 1995, pp. 30-31.

is to say, either we are going to pick out an aspiration level to meet, or we are going to make use of agents who set the aspiration level. Either general project has many distinct variations. One objective aspiration level could be imposed on all agents, or each separate agent could have their own aspiration level imposed upon them. Subjective aspiration levels could be obtained from each agent, or they could vote on a common aspiration level. All of these projects fail to produce a meaningful result, for quite general reasons.

Objectively defined aspiration levels fail to give aid to the contractarian enterprise for the trivial reason that they are not subjective. Separating the result from the preferences and beliefs of each person separates it simultaneously from any authoritative force derived therefrom. Any attempt to define an aspiration level subjectively fails for the very reason that we wished to introduce the satisficing agent into this effort: any effort to gauge whether a given set of rules meets a certain aspiration level depends on its expected utility for each agent being calculable. This in turn depends on each agent's preferences over outcomes and beliefs being known, as was made clear in chapter 2. It is exactly each agent's beliefs and preferences that are lacking. And so no satisficing agent that relies upon an aspiration level is going to be useful. It is also the case that no satisficing agent that relies on anything other than an aspiration level is going to be useful, as it would then identify a set of rules as acceptable for reasons other than a person's beliefs and desires, which would diverge from the instrumental spirit of this project. And so we conclude that satisficing agent will not, in and of themselves, be of any use to the contractarian.

IV: Conclusion

What then may we conclude from this investigation? The conclusion is obvious. Finding our previously proposed maximizers impossible to construct at the moment, we turned to survey the satisficing alternatives. These alternatives were ultimately found lacking because they either failed to be instrumentally valuable, or depended upon the very same inputs that we could not make use of when attempting to construct our maximizing agents. Given that the contractarian project is instrumental in nature, our only recourse is to focus upon the construction of agents which do not require an exhaustive enumeration of each person's beliefs and desires, while still retaining some strong relation to the same. We need agents that are not exhaustively informed, but whose recommendations would be found compelling. This will be the focus of our next two chapters. Thomas Hobbes and David Gauthier are the authors of arguably the two most famous attempts to carry out such a project. We will delineate each of their projects in turn, and investigate what may be said for and against each.

CHAPTER 4
~Hobbes' Universal Motions and *Leviathan*~

Thomas Hobbes is likely to remain the most famous instrumental contractarian for some time. *Leviathan* is also likely to remain his most read work. In *Leviathan,* Hobbes argues for the positing of a universal tendency to be found in all men that makes it reasonable to conclude that all men will eventually come into conflict with each other. Mankind will fall into an eternal struggle, each against the other, unless they all agree to institute a government of unlimited license to reign over all. We will ignore Hobbes' conclusions regarding the justified form and function of government, and focus on his argument for the positing of a universal tendency that may be ascribed to all persons. It is by focusing upon this tendency alone that we will most clearly see what Hobbes' efforts have to teach us about the properties that our required partial list of preferences must have.

In *Leviathan* Hobbes presents arguably the most complex and interesting effort to ground political obligation and the authority of morality in people's economic deliberation produced to date. Notwithstanding its sometimes primitive tools of analysis, *Leviathan* continues to capture theorists' interests and imagination.[98] This is in no small part due to his trademark theory of human nature. Notwithstanding the sheer volume of critical gazes brought to bear on this work, there is still significant divergence of opinion regarding what Hobbes' theory of human nature actually amounts to, and what part it plays in his contractarian endeavors. Given that it is our intention to examine what, exactly, Hobbes' theory of human nature is, and what we may glean from it, it is crucial

[98]We do not intend to be unduly harsh to Hobbes' analytic capacities. Ian Hacking points out that Hobbes adopts the frequency theory of probability in *The Emergence of Probability*, Cambridge University Press,

for us to be absolutely clear on these issues. It might be observed that the volume of critical analysis makes our task insurmountable in the space devoted to it, and this observation is not without merit. Perhaps the best that can be hoped for is to form a settled opinion regarding what it is that Hobbes' *Leviathan* presents for our consideration, albeit one that purports to be in large part accurate. We will therefore be engaging in a purely descriptive enterprise, thereby distancing ourselves from some recent contractarian investigators, whose (at least) partial aim has been to present a corrected version of Hobbes' arguments.[99]

Our purpose also allows another delineation of the field of study. We will not be pursuing the interconnection between Hobbes' theological and secular ratiocination. Our interest is purely secular. It may be objected that this will necessarily distort any interpretive efforts we thereafter make, and this may indeed be the case, but here this objection has no force.[100] This objection would be telling against efforts to either accurately portray Hobbes' rhetorical efforts *simpliciter*, or to assess Hobbes' failings as a theoretician. Our goal is neither of these. We are concerned with coming to an accurate portrayal of Hobbes' *as a contractarian*. We wish uncover how it is that the most famous advocate of this theory attempts to provide a compelling version of it, which nevertheless does not provide an exhaustive account of each person's psychology. We therefore attend almost exclusively to his descriptions of his method, and to his

---

1984, p. 48. In the 1600's this was subtle reasoning indeed. He did not, however, have the tools of game theory at his disposal.

[99] For example, the accounts found in Gregory Kavka's Hobbesian Moral and Political Theory, Princeton University Press, 1986 and Jean Hampton's Hobbes and the Social Contract Tradition, Cambridge University Press, 1986.

[100] This objection has been at least implicilty made by David Johnston in his *The Rhetoric of Leviathan*, Princeton University press, 1986. Johnston suggests that recent Hobbes scholarship focusing mostly on his political philosophy has effected a distorted view of *Leviathan*, one in which the political, theological, and

interpretation of mankind. We do not, therefore, require an exhaustive investigation of all of Hobbes' arguments, however they may be related; nor need we pass judgment on whether his arguments are valid or sound.

We limit our investigation to *Leviathan* due to two simple and widely acknowledged observations. The first is that Hobbes changes his arguments from book to book, and second that *Leviathan* is the superior presentation of his general doctrine.[101] One of the results of the first observation is that we will not spend a great deal of time examining Hobbes' other works, as there is no sufficiently important ambiguity in *Leviathan* which would require clarification by such a risky departure. Oddly, not all philosophers interested in an explicit recreation of Hobbes' arguments have drawn this inference. Gregory Kavka's *Hobbsian Moral and Political Theory*[102] quite explicitly recognizes that Hobbes' argument vary from text to text,[103] and nevertheless proceeds to make use of Hobbes' other works when analyzing *Leviathan*'s structure.[104] R. E. Ewin seems not to even notice that Hobbes' arguments and methods change over time when he describes Hobbes' method, freely making use of passages from *Leviathan*, *The Elements of Law*, *A Dialogue between a Philosopher and a Student of the Common Laws of*

---

physical trains of thought are percieved as almost unrelated. An accurate reading of *Leviathan*, he suggests, can result only if there is an effort to examine the text as a connected whole.

[101] While widely acknowledged, this latter opinion is by no means universal. Bernard Gert, for example, contends that *De Cive* is superior to *Leviathan* as a philosophical work in his introduction to "Man and Citizen", B. Gert, ed., Hackett press, 1993, p. 3. As Gert recognizes, however, *De Cive* lacks an incorporated account of human nature and as a result is unsuitable for our purposes. That this was to be supplemented by the account of human nature in *De Homine* is not sufficient for us to begin examining two texts for one coherent position when there is an obvious alternative available.

[102] Gregory Kavka, *Hobbesian Moral and Political Theory*, Princeton University Press, 1986. Kavka proposes to give an accurate picture of *Leviathan* on p. xiv of his preface.

[103] Ibid, p. 87.

[104] Ibid, p. 97, 157.

England, *De Cive*, *De Homine*, and *De Corpore* while making no case for this unitary interpretation of Hobbes' method being the correct one.[105]

Finally, given our purposes, we will assume that *Leviathan* is what it appears to be: a contractarian enterprise. Starting with agents attempting to maximize their expected utility, *Leviathan* sets out to prove that it is rational for each such agent to join together with all others to form a state, given the available alternatives. Any suggestion that Hobbes presented a strictly deontological theory, or that his theory is not based mainly on prudential considerations, or any other such other suitably radical proposal is surely "absurd on the face of it, and absurd on reflection."[106]

I: The Project

There is some disagreement over the form of Hobbes' enterprise. Hobbes recognizes quite clearly the differences between an empirical investigation and a deductive argument. He takes the former to be concerned with facts about the world, from which no general conclusions may be derived. The latter concerns what may always be rightly be said to follow from particular premises, yet has no direct application to the world. While some passages in Hobbes quite clearly lead one to believe he is engaging in a deductive enterprise, others quite clearly involve empirical evidence. This tension has led to a corresponding tension in the secondary literature. Some take Hobbes at his word, interpreting Hobbes as presenting us with definitions from which he presents analytic arguments from which he derives his conclusions, while others think that

---

[105] Notwithstanding this puzzling feature, the presentation of Hobbes' method is both revealing and interesting. See R. E. Ewin, *Virtues and Rights: The Moral Philosophy of Thomas Hobbes*, Westview Press, 1991, esp. the introduction and chapter one.

[106] David Gauthier, *Logic of Leviathan*, Oxford at the Clarendon Press, 1969, p. 28.

Hobbes' reliance upon empirical evidence makes such an interpretation implausible, notwithstanding Hobbes express thoughts on the matter.[107] Turning instead to passages where Hobbes suggests that internal investigation of one's own mind, and the observation of others are the only ways in which to demonstrate the truth of his conclusions, they conclude that he must have had some other method in mind when constructing the arguments of *Leviathan*.[108]

We will here argue that regarding *Leviathan* as Hobbes suggests, i.e. as the presentation of a deductive argument, or what he calles a scientific enterprise, makes sense of both his expressed opinions of what he is attempting as well as his appeals to empirical evidence. This interpretation further makes sense of Hobbes' claims that *Leviathan's* conclusions are certain, while human reason does not, in general, produce certain conclusions. Hobbes believes that only a public presentation of a derivation dependent upon definitions, which perspicuously convinces others of its truth, is a sign of certain science. This is explicitly distanced from mere Prudence, which is derived solely by experience, and whose conclusions are uncertain.[109] There must, by implication, be more to Hobbes' scientific presentation in *Leviathan* than experiential observations. To adequately identify Hobbes' method we will focus on chapter 5 of *Leviathan*, where he presents these differences most clearly. Correctly identifying Hobbes' method in

---

[107] See F. S. McNeilly's *The Anatomy of Leviathan*, Macmillan Press, 1968, for an example of the former. Kavka's *Hobbesian Moral and Political Theory*, provides an example of the latter. Kavka also denies that Hobbes' arguments could be deductive given his use of probabilistic reasoning. It will become clear that this observation does not count against the deductive interpretation of *Leviathan's* structure. The probabilistic reasoning Kavka points to is embedded in a deductive argument, and so it is best understood as a deductive conclusion. The statement that "These agents will, of necessity, decide that it is most probably the case that they will be attacked, and so will fortify themselves" is clearly not probabilistic in any sense that renders its inclusion in a deductive argument problematic.

[108] The most obvious passage in *Leviathan* that lends itself to such an interpretation reads "when I shall have set down my own reading orderly, and perspicuously, the pains left another, will be onely to consider, if he also find not the same in himself. For this kind of Doctrine, admitteth no other Demonstration.", p. 83.

Leviathan is crucial for our enterprise; an inaccurate description of *Leviathan*'s structure cannot be reasonably supposed to result in an accurate assessment of the applicability *Leviathan's* structure to our project.

I.i: Reason and Science in *Leviathan*

Hobbes, uncontroversially, considers language to be the most useful creation mankind possesses. It allows us to register our thoughts, and thus more easily recall the consequences of our thoughts, as well as to convey these thoughts to another. It also, through the use of general names, allows us to conceive, and utter, universal statements. Not an unmixed blessing, it can also allow us to incorrectly attempt the same. We can misregister our thoughts, by using words in an inconsistent manner, and in effect mislead ourselves regarding what it is that we have registered. We can also deceive others, either intentionally or not. And we can incorrectly reason about general terms, which leads to absurdity. Recognizing all this, Hobbes must assure the reader that his conclusions have not fallen into absurdity, and in so doing, it will become clear that he cannot be appealing to experience alone.

> REASON ... is nothing but *reckoning* ... of the Consequences of generall names agreed upon, for the *marking* and *signifying* of our thoughts; I say *marking* them, when we reckon by our selves; and *signifying*, when we demonstrate, or approve our reckonings to other men.[110]
>
> Not but that Reason it selfe is always Right Reason, as well as Arithmetique is a certain and infallible Art: But no one mans Reason, nor the Reason of any one number of men, makes the certaintie..."[111]
>
> The Use and End of Reason, is not the finding of the summe, and truth of one, or a few consequences, remote from the first definitions, and settled

---

[109] Leviathan, p. 115 and 97 respectively.
[110] *Leviathan*, p. 111.
[111] *Leviathan*, p. 111.

> significations of the names; but to begin at these; and proceed from one consequence to another...[112]

In the above quotations we find our puzzle presented most starkly. How are we to be assured of the truth of the consequences of any reasoning when no number of people reasoning in a like fashion can assure us of the same? Hobbes' reply to this question could be that he only meant to suggest that it is no number of people agreeing about the conclusions of any reckoning *alone* that guarantees the avoidance of absurdity, but it is instead any person (or any number of people) reasoning *in the right way* that provides the guarantee. Of course, such a response is useless without a sign whereby one may distinguish correct reasoning from that which is faulty or uncertain (absurd). Hobbes' sign of right reason (science) is the ability to perspicuously demonstrate the validity of one's conclusions to another.

When reasoning linguistically, be it internally or externally, people may be led astray for want of method. Hobbes suggests two main causes of this lack of methodical progress: failing to begin reasoning from definitions to conclusions via syllogisms, and failing to clearly and unambiguously define one's terms. He suggests that people mainly fail to follow proper method from stubbornness in wishing to maintain their opinion, and from unthinking acceptance of customary definitions or conclusions. Whatever the reason, it is clear that for any reason to be right reason, clear and unambiguous definitions must originally be made use of. This becomes entirely obvious when Hobbes presents us with the causes of reasoning to absurd conclusions, the first cause of which is beginning to reason without definitions. Such efforts are seen to be as absurd as attempting to add or subtract without knowing what "one", "two", or "three" signifies.

---

[112] *Leviathan*, 112.

The remaining causes of absurdity may be summarized as the misapplication of terms. To give the names of bodies to accidents, or vice versa, or making use of names that do not refer, or are meaningless, are all possible causes of absurd conclusions. To him that can avoid these pitfalls, "it is not easie to fall into any absurdity, unlesse it be by the length of an account...For all men by nature reason alike, and well, when they have good principles."[113] And so if we follow the method of Hobbes, we should all, reasoning from the same definitions, and making use of the same syllogisms, come to the same conclusions.

It is obvious that if this account is correct, reasoning is hypothetical in nature. Reasoning from definitions to conclusions does not thereby apply to the real world. Any reasoning about a triangle's properties, or of three-dimensional space, does not tell us anything about the world unless it is accepted that *this* thing is a triangle, or that we exist in three-dimensional space. If Hobbes is seen to recognize this implication, then so much the better for our account. Unfortunately, immediately proceeding his conclusions regarding how to avoid absurdity, Hobbes presents us with a rather puzzling passage:

> And whereas Sense and Memory are but knowledge of Fact, which is a thing past, and irrevocable; *Science* is the knowledge of Consequences, and dependance of one fact upon another...[114]

This seems to suggest either that science is not right reason, or that right reasoning involves matters of fact. Given that the science of which he most commonly speaks is geometry, and that he identifies science with completed right reason when he asserts that "we come to a knowledge of all the Consequences of names appertaining to the subject in

---

[113] *Leviathan*, p. 115.
[114] *Leviathan*, p. 115.

hand; and that is it, men call SCIENCE",[115] we seem to be left with the suggestion that complete reasoning (science) involves empirical evidence. Hobbes does not define "fact" in *Leviathan*, and so we are left with the task of sifting through alternate uses of the word in order to see what he might here mean. Unfortunately, Hobbes does not use the term in an altogether constant manner. He sometimes uses it to indicate a subjective opinion ("...to make him condemn some fact of his own..."[116]) and sometimes to indicate empirical data (as he seems to above.) And so we turn to alternative passages in *Leviathan*, in which Hobbes distinguishes knowledge of fact and science, in order to clear up this confusion.

> No discourse whatsoever, can End in absolute knowledge of Fact, past, or to come. For as for the knowledge of Fact, it is originally, Sense; and ever after, Memory. And for the knowledge of Consequence, which I have said before is called Science, it is not Absolute, but Conditionall.[117]

> There are of KNOWLEDGE two kinds; whereof one is *Knowledge of Fact*: the other *Knowledge of the Consequence of one Affirmation to another*... The later is called *Science*; and is *Conditionall*; as when we know, that, *If the figure showne be a circle, then any straight line through the Center shall divide it into two equall parts*. And this is the Knowledge required in a Philosopher, that is to say, of him that pretends to Reasoning[118]

These passages are as clear as one might hope for. Hobbes thinks that science, and right reason, bring about certain conclusions, that nevertheless are conditional. They do not straight-away tell us anything about the world, and any application to the world does not result in certainty. We may dismiss the original passage above by suggesting that, notwithstanding all the literary mastery apparent in *Leviathan*, Hobbes simply hit upon an

---

[115] *Leviathan*, p. 115.
[116] *Leviathan*, p. 124.
[117] *Leviathan*, p. 131.
[118] *Leviathan*, pp. 147-148.

unfortunate phrase. We are able to distinguish right reasoning from wrong reasoning, and it is to this to which we know turn.

It is quite clear to Hobbes that not all efforts at scientific investigation produce certain conclusions. Some efforts, however, do present certain conclusions. Better still, Hobbes supposed that we can distinguish between them. Given that we all reason in the same way (and well), a deduction that proceeds from clear, unambiguous definitions will convince others when presented publicly.[119] And so he concludes that "The signes of Science, are ... Certain, when he that pretendeth the Science of any thing, can teach the same; that is to say, demonstrate the truth thereof perspicuously to another..."[120]

And so we may conclude that *if* Hobbes presented a scientific account in *Leviathan*, *then* that would make sense of his insistence that his conclusions were certain. We rule out other possible methods of attaining certain conclusions because Hobbes does not discuss any.[121] We thus have some grounds for concluding that Hobbes must have been engaging in scientific reasoning. There is also other evidence that suggests that he was engaging in deductive reasoning. Hobbes claims that philosophers should engage in scientific argument.[122] The structure of *Leviathan* closely resembles Hobbes' proposal for a science: proceeding first with many definitions, and continuing to draw out the implications of them. It also makes sense of the fact that he was also trying to convince people of the applicability of *Leviathan* to the real world. It is this that makes sense of his appeals to empirical evidence, such as when he suggests that people who wish to deny that his definition of man applies to the real world should see if they do not agree with it

---

[119] Hobbes left it unsaid that the person learning the demonstration would have to have no contrary or biased agenda of his or her own.
[120] *Leviathan*, p. 117.
[121] Sapience is also infallible, but it is defined as a gathering of "much science". *Leviathan*, p. 117.

in action, by arming themselves when going traveling, or locking their doors.[123] The purely empirical account, by contrast, cannot explain Hobbes' claims that his conclusions are certain.[124] And finally, the deductive account can make sense of the passage made much of by those who wish to deny the deductive account, to wit:

> yet, when I shall have set down my own reading orderly, and perspicuously, the pains left another, will be onely to consider, if he also find not the same in himself. For this kind of Doctrine, admitteth no other Demonstration.[125]

Given that his reference to an orderly and perspicuous presentation obviously implies the presentation of a scientific account, it becomes clear that Hobbes is presenting a request of his reader to check his proof, and see if the deduction is correct. Hobbes requests that the reader investigate whether the same results follow from the reader's use of right reason.

So we conclude that Hobbes is here presenting a deductive account, not merely an empirical one. This interpretation makes sense of both the passages where Hobbes says he engages in deductive reasoning, and the passages appealing to facts. This also serves our interests as well. Were Hobbes to be found to be simply stating facts about his time, these results would be of much less interest for our purposes. Finding the exact details of our own time lacking, we have turned to Hobbes to see what alternatives he provides for contractarian accounts in general. And it is with this in mind, as always, that we proceed to his account of human nature.

---

[122] *Leviathan*, pp. 147-148.
[123] *Leviathan*, pp. 186-187.
[124] Without resorting to an incredibly uncharitable interpretation of Hobbes' integrity as a scholar.
[125] *Leviathan*, p. 83.

I.ii: Natural Man

Truly appreciating Hobbes' theory of human nature requires recalling his basic physical commitment: a body in motion stays in motion. It also requires a sympathetic imagination. Hobbes does not produce an exhaustive account of the physical components of a human being. Neither does he exhaustively examine our psychology - saying nothing, for instance, about the various ways in which we can combine our sensory experiences. This is not a failing in Hobbes' account. Hobbes is interested in identifying, through definition, various characteristics that we will attribute to actual people. Deriving what he can from these clear definitions, his hope is that we will confirm that people do possess these characteristics and thereby conclude that his political conclusions are relevant to the real world. This does not require an exhaustive account, so much as a presentation of definitions and a train of thought that his readers will find clear, and in which they will find a correct characterization of mankind.[126]

The physical properties of man being of interest only much later on in *Leviathan*, Hobbes begins by focusing primarily on mental properties. All thoughts originate from sense experience. The motions of the external world put pressure on our sense organs, which in turn produce fancies in our mind, and it is those immediate fancies that are called sense. As our sense organs are bodies, and we sense via movement of our organs, and a body in motion stays in motion, the immediate fancies must remain in our system after the original object that produced the original motion has moved on. This remaining

---

[126] Hobbes does not follow this procedure as explicitly as one might wish. This fact probably accounts for the disagreement regarding his method discussed above. Hobbes tends to present his theory not as a hypothetical construction, but as an actual examination of what man consists of. We will follow this style for its aesthetic qualities.

motion is what we call imagination, which Hobbes describes as "decaying sense".[127] He immediately points out that this ought not to be taken to imply that he thinks that imagination eventually slows down and ceases, contra his physics, but that each motion is eventually overcome by other motions, as "the light of the Sun obscureth the light of the starres."[128] It should be unsurprising that Hobbes notes as salient the imagination that is brought about by words and the like (which he calls understanding,) given science's dependence upon it. This understanding can involve both understanding simply the will of the utterer (perhaps by only the tone of the voice, for example) and also the understanding of the concepts which these words are to signify.

People can also manipulate their imagination, both by combining various discrete images, and by having one thought (image) follow after another in a train.[129] These trains are of two sorts: guided and unguided. Unguided thought is thought not regulated by some end, while guided thought aims at some end identified by a passion.[130] These trains of regulated thoughts are themselves of two kinds: those that seek out the cause of an imagined effect (seeking), and those that attempt to discover all of the possible effects of an imagined cause (invention). These powers exhaust the natural "Discourse of the Mind."[131]

Imagination also guides our endeavors. Endeavors are defined as the "small beginnings of [voluntary] Motion, within the body of Man, before they appear in

---

[127] *Leviathan*, p. 88.
[128] *Leviathan*, p. 88.
[129] We will modernize spelling when not directly quoting *Leviathan*.
[130] None of what is said here should be mistaken to imply that people are at some liberty here to direct their thoughts some way rather than another. Hobbes was a determinist; he took it that 'freedom' amounted to no more than a lack of external constraints.
[131] *Leviathan*, p. 96.

walking, [etc.]".[132] These may either be towards the cause of the endeavor (an appetite) or away from that cause (an aversion). As there can be no thought of directed action without an imagined end, and no imagined end without an original sense experience, it is obvious to Hobbes that all voluntary motions must be originally caused by sense experiences. Moreover, Hobbes posits that all sense experience causes an endeavor. This is not obvious, but can be taken as implied when Hobbes defines a man's contempt for some thing by suggesting that this is caused by a man's *overcoming* of the action of the thing sensed.[133] To be neutral towards some object, other internal motions must have been pitted against the motions produced by the experience. All of this seems to suggest that we could have no natural appetites or aversions, but this is not so. Hobbes suggests a few natural, general appetites: food, excretion, and exoneration.[134] Interestingly, aversions need not proceed from experience, as Hobbes suggests that we have an aversion to things whose consequences we do not know.[135]

We are now in a position to make sense of a much-misunderstood passage in *Leviathan*, to wit:

> Nor can a man any more live, whose Desires are at an end, then he, whose Senses and Imaginations are at a stand. ... And therefore the voluntary actions, and inclinations of all men, tend, not only to the procuring, but also to the assuring of a contented life; and differ only in the way: which ariseth partly from the diversity of passions... and partly from the difference of the knowledge...each one has of the causes, which produce the effect desired.
>
> So that in the first place, I put forward a generall inclination of all mankind, a perpetuall and restlesse desire of Power after power, that

[132] *Leviathan*, p. 119.
[133] *Leviathan*, p. 120.
[134] *Leviathan*, pp. 119-120. Exoneration here refers to the evacuation of the bowels. This passage requires that we either: 1) conclude that these general appetites are not thoughts in the mind, or 2) conclude that these general appetites are to be understood as exceptions to Hobbes' preliminary remark that "The originall of ... all [thoughts] is that which we call SENSE..." [p. 85].
[135] *Leviathan*, p. 120.

> ceaseth onely in Death... because he cannot assure the power and means to live well, which he hath present, without the acquisition of more.[136]

It is not necessary to claim that Hobbes assumed, without so stating, that some men's desires are without limit in order to correctly derive the claim that there is a general inclination in men to desire (be drawn to gathering) ever increasing power. C. B. Macpherson is drawn to endorse this position by misunderstanding the strength of Hobbes' claim, together with his perhaps assuming that desires result in action.[137] Hobbes is not here claiming that it is the case that each person *actually* strives for power after power, but instead is simply saying that given human nature there is reason to conclude that each person has that *tendency*. Tendencies, as motions, may yet be overcome, and do not themselves result in action or manifest themselves as expressions of the will.

If Macpherson's interpretation is to be believed, then this passage renders a later passage, in which Hobbes famously proves that the state of nature is a state of war, entirely irrelevant.[138] Given scarcity we could forthwith conclude that people in a state of nature are in a state of war. Nevertheless, Hobbes does not do so. He instead suggests that in the state of nature, given rough equality, we each can anticipate that it is likely that others will come to desire the same goods, and that therefore we will be enemies, and that enemies seek to destroy each other, and that this leads to anticipatory strikes and that

---

[136] *Leviathan*, p. 161. We have excluded the account of felicity in the quotation because it serves only to obscure the issue, and there is little point in examining a term on which Hobbes himself was so evidently confused. See McNielly's *The Anatomy of Leviathan* (esp. pp. 129-136) for an excellent discussion of how confused and vacuous Hobbes' discussions of felicity are in *Leviathan*. We have also excluded the inclusion of stated causes of the desire for power which do not always lead to it; they are supposed to be included merely for completeness, and therefore they too only obscure the passage.

[137] See his introduction to *Leviathan*, esp. pp. 32-37.

[138] *Leviathan*, p. 184.

therefore the state of nature is a state of war.[139] Any interpretation that claims that Hobbes' conclusion here is so strong as Macpherson's does must contend with this same objection. We claim instead that the first passage is designed to make it reasonable for people to anticipate that others will become interested in acquiring the same goods. We now turn to how he shows this.

Given that desires are positive endeavors, which are caused by imaginations,[140] which are in turn caused by the senses, Hobbes' conclusion - "Nor can a man any more live, whose Desires are at an end, than he, whose Senses and Imaginations are at a stand"[141] - is almost analytic. As long as one has sense, one will likely have desires. It would have been completely analytic if Hobbes had included having contempt and aversion for things sensed as well, but we cannot expect the life of the apologist to be trouble-free.[142]

Turning now to the tendency towards procuring and assuring a contented life, we suggest that the key lies in recalling the two trains of regulated thought, wherein we examine each imagination for its cause, and for all possible effects. As our senses will always be affected by the world, we will, insofar as we are directing our thoughts, examine each imagination for its causes, and its effects. And upon desiring the supposed cause of this imagination, or upon discovering it to be valuable or useful, either now or in

---

[139] Discounting, for the sake of simplicity, the secondary cause of glory.
[140] We use the more awkward term "imaginations" instead of "images" so as to avoid implying that all sense is visual.
[141] *Leviathan*, p. 160.
[142] One can avoid this trouble if one supposes that, contra his explicit definition, "desire' here refers to any endeavor whatever, as he seems to when he says "...as Appetite of food, Appetite of excretion, and exoneration (which may also and more properly be called Aversions, frome somewhat they feele in their Bodies..." [p. 120] However this evidence does not seem decisive.

the future,[143] we will attempt to acquire it, or find a way through which it may be acquired.

Given our natural reactions to sense data, we will always be examining things, and come to the conclusion, should these things be power, that we wish to obtain them. Since power is only a thing that helps us secure the object of a desire, this should be unsurprising. That we will always seek to determine if this or that thing is a power (could be useful to us) is guaranteed by our basic thought processes. That we cannot secure our present means without the acquisition of more power is derived straightforwardly from the fact that prudential reasoning is not science, and therefore not certain. We can never be assured that we have completely secured our present goods, and so will always conclude that there is more that we can do to further secure them.

Given that we cannot be certain about our efforts to secure such goods as we now possess, our minds naturally tend to shape our voluntary motions in such a way as to continue to gather new resources (new powers), albeit only so far as we tend to engage in directed thoughts. When engaging in scientific inquiry regarding man in the state of nature, it is this conclusion about our basic human nature that allows us to reckon it likely that people will tend to attempt to gain possession of the same things. This, as has been previously mentioned, allows one to eventually conclude that the state of nature is a state of war. But this brings us beyond that part of Hobbes' project that is of interest to us here, and it is to what can be said both for and against Hobbes' method that we now turn.

[143] Hobbes explicitly defines three types of good which may be attributed to a thing, good in the future, good in itself, and good as a means.

II: Critiquing Hobbes' Method

In Hobbes' eyes, *Leviathan*'s most obvious strength is that it is a deductive, not empirical, effort. That is to say that, in his own terms, Hobbes attempted to derive his contractarian conclusions purely from a clear and unambiguous definition of what it is to be human, rather than an effort to examine the actions of people during his day, and derive his conclusions from such observations. This is not to say that he paid no attention to the people around him, as he must have had some cause to hope that his readers would recognize that his account of human nature accurately portrayed humanity. Rather, he engaged in an effort to render conclusions that were going to be valid for all people, for all time.

Notwithstanding the issue of whether or not his conclusions follow from his premises, Hobbes' deductive account fails to provide sufficient assurance of the relevance of his conclusions to any particular population of people. The problem depends on Hobbes concluding that there is merely a tendency towards power after power, and that it is this alone that makes it reasonable to anticipate the state of war, which in turn justifies his conclusions. More specifically, it is not the content of this tendency that is objectionable but the reliance on a tendency itself that renders the effort less than satisfactory.

Observing a mere tendency that universally obtains is not, of itself, a strong enough observation to conclude that all people will be so motivated by that tendency in any particular population. In other words, while it may be the case that this tendency will be taken into account by anyone who is aware of it while they attempt to predict other's

behaviors, it is not the case that this tendency will in general, or ever, result in overt action.

In a critique of *Leviathan*, this observation may lead one to suggest, for example, that while accepting that all people have this tendency to endeavor towards power after power, each person's learned desires tend towards resting and hiding.[144] Moreover, one could claim that the motions begun by these desires are stronger than those behind the drive for power after power, so that it is *unreasonable* to conclude that the state of nature is the state of war.

For our purposes, this translates into the concern that whatever tendency is relied upon, however basic a part of human nature it is, it is not yet guaranteed that the people for whom the contractarian calculation is being run would act on this tendency. Substituting a basic, or universal, desire for Hobbes' basic tendency does not change the picture significantly. However basic this desire is, it is not yet guaranteed that people care about it sufficiently to endorse a set of rules based upon it. For a compelling account to emerge, it is required that the desire used is sufficiently strong, or is held to be sufficiently important, that competing desires are extremely unlikely to be held to be more important. Indeed, it is this strength alone that is required. Universally held desires are a tidy tool with which to produce a wieldly theory, but they are, strictly speaking, unnecessary. A contractarian account can be produced in which different members of the population consider entirely different things important, as long as the contractarian identifies the preferences that are of paramount import to each type of agent. Gauthier takes the latter part of this observation to heart, and identifying market transactions as

[144] This example is inspired by Kavka's discussion of the rationality of hiding out in the state of nature, and hoping that no one finds you.

those which we may suppose are of paramount import, proceeds to construct his market contractarianism in an effort to present a fundamental justification of his moral theory in *Morals by Agreement*. It is an examination of this effort which will occupy us in Chapter 5.

Chapter 5
~ Gauthier's Market Contractarianism ~

In *Morals by Agreement*, Gauthier attempts to guide his theory between two almost contradictory observations, given his allegiance to a rational choice theoretic conception of reason and a commitment to reconciling rationality and morality. On the one hand, he recognizes that if a moral theory were to produce duties genuinely in the interest of each individual agent, "morality would be superfluous."[145] Rational agents have no need of counsel instructing them that they must act in their best interest – they do so naturally. On the other hand, no moral theory which cannot recommend itself to an agent's rationality has any hope of being effective in action. Given a commitment to instrumental rationality, where to act rationally is to act in an effort to maximize the fulfillment of one's interests, it seems inevitable in light of these observations that the construction of a rationally compelling moral theory is doomed to fail. What is a contractarian to do?

These observations seem to allow for no navigation between them. In offering his solution to this dilemma, Gauthier proposes a heterodox account of rational choice theory, introducing the notion of a rational *disposition* by which to avoid falling prey to the consequences of the first observation while respecting the second. It will be recalled that we have embraced the orthodox rational choice theoretic account. It will also be recalled that we have not, to this point, made any argument for the rejection of Gauthier's account of rational choice. In this chapter we justify our allegiance to the orthodoxy. In doing so it must be clear that we are able to avoid the dilemma posed above, and still

[145] David Gauthier, *Morals by Agreement*, Oxford University Press, 1986, p. 1

produce a contractarian morality. And so the flavor of this chapter will, of necessity, occasionally adopt a more critical tone than that of the previous chapter.

We should not be understood to be diverging from the reasons that brought us, in the previous chapter, to examine *Leviathan* with such care, so much as supplementing them with a critical analysis. It is still our intention to search Gauthier's work for what it may tell us about our efforts to construct a non-exhaustive account of peoples' interests and beliefs which, when applied to a contractarian endeavor, nevertheless will result in a maximally compelling justification. We still suggest focusing on what is widely acknowledged as Gauthier's most masterful and compelling presentation of his efforts (*Morals by Agreement*). And we will still be more interested in examining the logical structure of Gauthier's account than in examining whether his output is compelling as a recognizably moral theory.[146]

The one stylistic different between our presentation of Hobbes' efforts and Gauthier's is that we will be less focused on textual evidence and the clarification of the author's intent, preferring instead to thematically develop and summarize Gauthier's commitments. This luxury is afforded us for a number of reasons. Due to stylistic developments in the three and a third centuries between *Leviathan* and *Morals by Agreement*, Gauthier is significantly clearer when identifying his intentions and arguments than Hobbes was. Further, Gauthier's prose is also less troublesome to interpret, as his style is more familiar to us than Hobbes'. Finally there are relatively fewer interpretive traditions that fundamentally misunderstand *Morals by Agreement* than *Leviathan*; and so we have less to overcome. This is also a necessity forced upon us by

[146] Accordingly we will concentrate on presenting the structure of chapters IV through VII.

the complexity of Gauthier's arguments. To accurately present a detailed analysis of Gauthier's arguments would be well beyond the scope of the present effort.

I: The Project

In an effort to prove the rational obligatoriness of morality, Gauthier must, analytically, provide all people reason to adopt and adhere to the content of morality. In doing so he must first provide some goal to which all strive, thereby providing the motivation to so adhere. Some state of affairs, superior to our natural condition, towards which we find ourselves willing to proceed over all other possible options, must be presented. To fill this role, Gauthier presents the perfectly competitive market. It is the perfectly competitive market in which Gauthier finds sufficient promise of mutual advantage, should it be realized, to argue that our rational reflection would counsel attempts to emulate the same. The conditions that define a perfectly competitive market are thought superior in the same way that any defined market is superior: transactions are possible only within a market, and transactions are, by definition, desired by the participants. All participants, therefore, desire a market. It is still a question whether or not all persons wish to be participants, something we shall address below.

Simultaneously, in the market Gauthier finds patterns of behavior in which the dictates of morality would have no meaningful place. In the perfectly competitive market each person acting so as to maximize his or her utility tends towards a distribution of goods that is in equilibrium; each will eventually prefer what he or she owns through trade to any other bundle of goods available. This state of affairs is also found to be Pareto optimal; agents acting from self-interest produce a state of affairs in which no one could be made better off without someone else being made worse off. In a perfectly

competitive market there arise no cases where one person attending to self-interest leaves some other worse off than the latter would otherwise be. As Gauthier identifies these cases as those to which moral constraints correctly apply, he claims that the perfectly competitive market is a *morally free zone*. For Gauthier, morality is relevant only in market situations that are not perfectly competitive. Morality finds its *reason d'etre* to be the correction of market failures, and that exclusively.

In the perfectly competitive market, people act under certainty[147] in an effort to maximize their consumption of products while minimizing their provision of factor services in both production and exchange. Not only certain about both the fixed nature of their circumstances and the characteristics of the same, actors are also certain of the actions and reactions of their fellows, and so make choices parametrically. All products and factors of production are assumed to be exclusively and exhaustively owned. That is, everything is owned; and each good is owned by only one individual.

All goods are further supposed to affect only one person's utility function, and so consumption is exclusive, as is enjoyment. Once a good is consumed, no other person may consume it, and "each person's utility is strictly determined by the goods he consumes and the factor services he provides."[148] This latter stipulation is already familiar to the reader as Wicksteed's assumption of non-tuism, discussed in chapter 2. These assumptions are all necessary to guarantee the absence of externalities – effects on some person's utility arising from an act of production, exchange, or consumption where this person is not a participant, or not a willing participant, in the exchange. The existence of externalities upsets the matching of supply and demand, which in turn may

[147] Certainty may be contrasted to the uncertainty outlined in Chapter 2 of this thesis.
[148] *Morals By Agreement*, p. 86.

upset the equilibrium outcome of the market being optimal. The stipulations that define a perfectly competitive market are jointly sufficient to ensure that the market equilibrium is also optimal.[149]

Appealing to the intuition that Robinson Crusoe has no complaint to make about the outcome of his choices on a deserted island, Gauthier argues that each individual has no complaint to make for her treatment in a perfectly competitive market. Just as each of Robinson Crusoe's choices was completely voluntary, so too is each person's choice (and outcome) when situated in a perfectly competitive market. The results being known with certainty, and chosen parametrically, no one may claim that the workings of the market treated him or her unfairly. Each person's gain in market activity is equal to the worth of her voluntary contribution. No one is singled out for special treatment – preferential or otherwise. Where equilibrium and optimum coincide in the perfectly competitive market, there is no place for morality as impartial constraint. Morality has no place because each chooses freely to be affected by any transaction, the benefits of which are proportional to her contribution; and the optimality of the market means that any divergence would make one person better off only at the expense of some other – an expense that this unfortunate other would not agree to, given the assumption of non-tuism. Any forced deviation from the workings of a perfectly competitive market would therefore involve choosing to value one person's advantage over another person's corresponding disadvantage – to treat people partially.

As Gauthier recognizes, this result applies only to the workings of the market, not to the initial distribution of factor endowments. "But neither the operation of the market

[149] This analysis does not include a discussion of rent, as explicitly recognized by Gauthier in *Morals by Agreement*, p. 98.

nor its outcome can show, or can even tend to show, that its initial situation is ... either rationally or morally acceptable."[150] Fair outcomes are the results of the perfect market working from a fair initial distribution. Modifying the unacceptable Hobbesian claim that each person's initial endowment is whatever they can make use of, Gauthier presents an alternative which is more compatible with the perfectly competitive market. Making use of Hobbes' description of initial endowments allows for non-private goods. If you are stronger than others, you then have a right to use their body, as do they. Gauthier proposes instead that each person's basic endowment is what he or she can make use of, and which no one else could make use of in his or her absence. This is easily seen to include each person's mental and physical abilities, but does seem to leave a lot of the world unclaimed. That this is the correct (rational and fair) conception of the basic endowment is the focus of Gauthier's arguments for his version of the Lockean proviso, to which we now turn.

Gauthier proposes that his Lockean proviso (hereafter simply 'the proviso') is the rational and fair way of extending each person's basic endowment in natural interaction (pre-agreements) to include those goods not yet so apportioned. He suggests for consideration that the proviso be understood as excluding the worsening of the situation of any person by predation or any other way, excepting only situations in which this is necessary to avoid worsening your own position. It is worthy of note that 'worsening' is here understood as a different animal from 'failing to make better off.'[151] As this proviso is intended to apply to interactions in a less comprehensive (and more intuitive) sense than that defined in chapter 2, Gauthier claims that "the proviso prohibits bettering one's

---

[150] *Morals by Agreement*, p. 94.
[151] See esp. *Morals by Agreement*, p. 204.

situation through interaction that worsens the situation of another. This, we claim, expresses the underlying idea of not taking advantage."[152] If this proviso is found to be both rational and fair, then it justifies not only the acquisition of property, but also treating personal abilities as being a part of each person's basic endowment. There is no way to make use of someone's body or mind against their will which would not run counter to this proviso.

That this is the rational baseline from which to enact any bargains comes from Gauthier's much-maligned supposition that bargains, which involve baselines derived at least in part from predatory activities, or the taking of advantage more generally, are not stable.[153] The reason for this is that it is not rational for one to be disposed to fulfill bargains in which predation helped determine the baseline. It is not rational to be so disposed because being so disposed would then encourage predators to take advantage of you. It would be better, overall, if you were not so disposed, and so did not provide an incentive for predation. If it is not rational to be so disposed, then it is not rational to comply with bargains thus made.[154] This, in turn, allows Gauthier to suggest that correcting all effects of taking advantage of potential cooperative partners in the past is a precondition for rational agreement. Without such a correction, no bargain is recommended, as no compliance is expected. The conclusion drawn is that any violation

---

[152] *Morals by Agreement*, p. 205.

[153] See for examples Jan Narveson "Gauthier on Distributive Justice and the Natural Baseline", *Contractarianism and Rational Choice*, Peter Vallentyne, ed., (New York: Cambridge University Press), 1991. pp. 127-148; Jean Hampton "Two Faces of Contractarian thought" *Contractarianism and Rational Choice*, Peter Vallentyne, ed., (New York: Cambridge University Press), 1991. pp. 31-55; Chris Tucker, "A Moral Obligation to Obey the State", *Journal of Value Inquiry*, Vol. 34:2-3, 2000, pp. 333-347.

[154] This depends in large part on Gauthier's heterodox rational choice theory, in which 'disposition' is a technical term. We will dwell upon it in greater detail below, however, and even an intuitive understanding of the term suffices for the level of discussion presented here.

of the proviso must be corrected, as the proviso forbids the taking of advantage. And so the rationality of the proviso as a pre-condition for rational bargaining is demonstrated.

That the proviso is fair is much less adequately defended – relying again on the identification of fairness with impartiality, and suggesting that to re-distribute any goods acquired in compliance with the proviso would require bettering one person at the expense of another. In other words, the redistribution would require partial treatment. That this may be suggested of any other exhaustive distribution of goods seems to trivialize Gauthier's argument, but it is not our intent to be overly critical, and we will waste no more time upon this point. Having identified Gauthier's rational pre-condition against which bargaining may take place, we turn to his argument that it is rational to dispose oneself to constrain pursuit of maximum expected utility in certain cooperative scenarios - arguments for the adoption of a moral disposition.[155]

Gauthier envisions a scenario in which a rational agent is attempting to determine which disposition she ought to adopt. Whether, on the one hand, to adopt a disposition to always act in such a way so as to maximize her own utility given the strategies she expects others to adopt in strategic situations (become a straightforward maximizer), or, on the other hand, to adopt the disposition of a constrained maximizer. The constrained maximizing disposition is best described (unsurprisingly) by Gauthier himself. In laying out the formal conditions of this disposition, Gauthier states that

> We shall therefore identify a constrained maximizer thus: (i) someone who is conditionally disposed to base her actions on a joint strategy or practice

[155] A significant part of the argument that the constraint argued for by Gauthier is moral constraint depends on the fact that the agreed distribution of the benefits of rational agreements will accord with his principle of minimax relative concession. His argument that this is both the rational agreement point, and that this point is recognizably moral, provides the linchpin to his argument that constrained maximizers adopt moral dispositions. We will not dwell upon Gauthier's development of minimax relative concession because it was never our intention to dwell overlong on considering whether his efforts adequately captured morality.

> should the utility she expects were everyone so to base his action be no less than what she would expect were everyone to employ individual strategies, and approach what she would expect from the co-operative outcome determined by minimax relative concession; (ii) someone who actually acts on this conditional disposition should her expected utility be greater than what she would expect were everyone to employ individual strategies. Or in other words, a *constrained maximizer is ready to co-operate in ways that, if followed by all, would yield outcomes that she would find beneficial and not unfair, and she does co-operate should she expect an actual practice or activity to be beneficial.* In determining the latter she must take into account that some persons will fail, or refuse, to act co-operatively.[156]

Choosing between these dispositions, a person need only consider the situations in which the two dispositions, if adopted, would counsel different courses of action. Given the definitions of straightforward and constrained maximization, above, these situations are characterized by the possibility of a fair and beneficial co-operative outcome, and also some advantage gained from defection via some deviation from the co-operative strategy.

In any interaction, constrained maximizers face either a straightforward maximizer or another constrained maximizer. If they face a constrained maximizer, then they are both in a position to realize the benefits of the co-operative outcome. If they face a straightforward maximizer, then they behave like a straightforward maximizer, and are not taken advantage of.

Straightforward maximizers also will face either a constrained maximizer or a straightforward maximizer. If they face another straightforward maximizer, then they both adopt individual strategies, and are not able to take advantage of the benefits of defecting from any co-operative strategy. If they face a constrained maximizer, then they still both adopt individual strategies, and both fail to take advantage of the benefits of defecting, but also fail to gain the benefits of adopting a co-operative strategy. No

---

[156] *Morals by Agreement*, p. 167, emphasis added.

benefit is realized from defection, yet constrained maximizers are able to take advantage of the benefits of joint strategies; and so the rational disposition to adopt is the disposition of constrained maximization. Thus is the rationality of moral constraint proved.

As Gauthier makes perfectly clear, the rationality of moral constraint depends on a heterodox rational choice theory. In navigating between the constraining nature of moral dictates and the interest driven account of rationality, Gauthier has created a gap between them by introducing the notion of a disposition. In Gauthier's proposed alternative, 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 disposition."[157] A choice, in turn, is rational if it is expressive of a rationally held disposition. In such a way can rationality maintain a link with interests, and nevertheless council constraint. Notwithstanding this substantial advantage, we do not support such a radical alternative conception of rationality. In what follows we suggest that this benefit may be had by the orthodox conception of rational choice, and so there is no need to adopt such a radical alternative. The cost of doing so is simply reinterpreting what it is that we suppose we are doing when we council that a person constrain their utility-maximizing activities. It is further suggested (although not much more than suggested) that dispositions would cause more trouble than they are worth.

II: Dispositions, Rationality, and Constraint

Gauthier suggests that a rational choice ought to be understood as one that expresses a disposition which it would be rational to adopt, and that a disposition is the rational one to adopt if and only if its adoption would maximize expected utility at least

[157] *Morals by Agreement*, p. 182-183.

as well as any other alternative. We say that such a radical departure from the standard account of rational choice is not necessary for Gauthier's purposes.[158] In proposing an account of rational morality that is nevertheless a constraint upon behavior, Gauthier conjures up an intermediate piece of mental machinery of which the rational agent may make use. In no other way can the link between the constraint of morality, and the interest based nature of rationality, be maintained. Gauthier claims that a rational disposition, chosen in an effort to maximize expected utility, can occasionally counsel one not to act in a straightforwardly maximizing manner. There are many problems with this account that extend far beyond any particulars of the argument – problems with the introduction of dispositions as rational apparatus. Dwelling on these problems would at this point be premature. If no alternative presents itself which can overcome the problems inherent in attempting to present a rationally compelling account of a moral theory that counsels constraining the pursuit of self interest, then, however problematic, the introduction of dispositions is a necessary evil. There is a viable alternative account available to us, however. We suggest that the contractarian project be understood as providing counsel as to which preferences it is rational to adopt in order to maximize the expected utility of one's current preferences.

It has become standard in the literature to distinguish between final ends and instrumental ends.[159] A final end is a goal that one has for no other reason. I prefer to do X because I think that X is a good thing to do. An instrumental end is an end that I have

---

[158] We will here be focusing on his contractarian purposes as described above. We will ignore Gauthier's work on the rationality of pre-commitment strategies in games of nuclear deterrence, as exemplified in Gauthier's "Deterrence, Maximization and Rationality", *Moral Dealing*, Cornell University press, 1990, pp. 298-321. These two issues are, however, closely linked.

[159] See, for example, David Schmidtz's *Rational Choice and Moral Agency*, Princeton University Press, 1995, esp. pp. 58-59. Schmidtz distinguishes between the commonly accepted final, instrumental and

acquired in order to further some other end that I have. I prefer to do X because I think that it will bring about Y, and I think that Y is a good thing. Y can be considered a good thing either instrumentally or finally. To take a timely (though of course counter-factual) example: the author of this thesis prefers to write this thesis for instrumental reasons; he believes that it is a way to bring it about that he acquires a Ph.D. in philosophy. Once he has internalized the project, the goal of writing a thesis is an instrumental goal. He had no previous interest in the writing of a thesis in and of itself – in fact it seemed like quite a bore! The end was only acquired in order to bring it about that the author would be likely to receive a Ph.D. in philosophy. It does not matter whether the goal of receiving the Ph.D. is itself an instrumental goal. It may be that the goal of receiving a Ph.D. in philosophy was adopted as a way to finally show that snotty grandfather of his how smart he really was, or it may not have been so adopted. The standard account of rational choice theory has no problem accounting for such a scenario, nor of evaluating the rationality of any choices made about the rationality of any given proposal. When choosing to adopt a new set of preferences from any given set of alternatives, the rational choice is to adopt the set of preferences which would maximize the expected utility of the set of preferences which the agent has at the time of the choice.

So too may the contractarian effort be interpreted within an orthodox rational choice theory. We can now roughly re-interpret Gauthier's results to show that given the choice between adopting no new preferences (the straightforward maximizer), and the adoption of a preference to co-operate in certain strategic situations given a certain class of potential partners, one should choose to adopt this new preference. Duncan MacIntosh

---

constitutive ends, and introduces a different kind of end: a maieutic. A maieutic end is an end to have other ends. We ignore this discussion for the sake of simplicity of presentation.

has proposed this alternative account while roundly criticizing Gauthier's efforts.[160] Since it is not yet our intention to criticize the adoption of dispositions, but only to suggest that orthodox rational choice theory can account for what it must, we will not dwell upon these critiques. It is worthwhile to note that MacIntosh has gone so far as to suggest a rather precise preference ordering which ought to be adopted. This list of preferences goes well beyond merely a re-interpretation of Gauthier's efforts, and is an effort to supplant both the method Gauthier proposes *and* the content. We have no desire to attempt the latter, and so will not present MacIntosh's account.[161] That Gauthier's arguments may be so re-interpreted may be adequately demonstrated by rephrasing his definition of the constrained maximizer, which we quoted above:

> "We shall therefore identify a constrained maximizer thus: (i) someone ~~who is conditionally disposed~~ [who prefers] to base her actions on a [fair] joint strategy or practice should the utility she expects were everyone so to base his action be no less than what she would expect were everyone to employ individual strategies... [given that]; (ii) ~~someone who actually acts on this conditional disposition should~~ her expected utility [also] be greater than what she would expect were everyone to employ individual strategies."[162]

In other words, a constrained maximizer may be thought of as someone who prefers to fairly co-operate with others when co-operative strategies afford some advantage over acting with individual strategies. The re-interpreted quotation above seems to repeat the second half of condition (i) in condition (ii). This is due to Gauthier's separation of

---

[160] See for examples Duncan MacIntosh, "Two Gauthier's?", *Dialogue*, 1991, pp. 3-32; "Preference-Revision and the Paradoxes of Instrumental Rationality" *Canadian Journal of Philosophy* (22:4), Dec. 1992, pp. 503-530; "Co-operative Solutions to the Prisoner's Dilemma", *Philosophical Studies*, (64:3) December 1991, pp. 309-321; "Retaliation Rationalized: Gauthier's Solution to the Deterrence Dilemma", *Pacific Philosophical Quarterly*, (72:1) March 1991, pp. 9-32.

[161] Those interested are referred to MacIntosh's "Co-operative Solutions to the Prisoner's Dilemma", *Philosophical Studies*, (64:3) December 1991, esp. pp. 316-317.

[162] *Morals by Agreement*, p. 167, strikethroughs have been added to indicate original text to be ignored and square bracketed text has been added.

action and disposition, so that this disposition must be described in such a way that it is clear that it is this disposition which governs action in these particular circumstances. There is no need for this description in the standard rational choice theoretic account, as preference leads directly to action.

And so while we have by no means ironed out all the problems for a contractarian proposal that envisions itself as recommending instrumentally valuable preferences, we have shown that such a proposal is *prima facie* possible.[163] This type of project does have the substantial advantage of being able to provide an answer to the infamous compliance problem. One follows through on the dictates contractarianism recommends because after adopting the instrumentally valuable preferences, you then desire to follow through on the dictates. There is one large problem for such an account, however. It may seem that by providing straightforward rational choice theoretic counsel, contractarianism fails to account for the role of constraint in moral theory. If there is no constraint in the contractarian account, then there can be no claim that the contractarian provides an account of morality.

Or so it may seem. Where Gauthier decided to create a wedge in order to separate moral constraint and rational choice with the introduction of dispositions, we choose to modify the presupposition of morality as requiring a kind of constraint. Given the intuitive force of the idea that morality must involve constraint - an intuition shared by Gauthier, Danielson, and common sense - we must attempt to neutralize that force before presenting an account of moral theory that tosses it aside. We turn to the examination of the intentions of a saint. Saints will here be identified with persons who have final ends

[163] One of the most glaring difficulties being that we have no argument at all for how strongly these preferences should be preferred in order to maximize expect utility.

that already conform to the dictate of a rational choice contractarian account. They will not need to constrain their actions to conform to the dictates of morality. If morality entails constraint, then they seem to not act morally. However this conclusion is at odds with our common judgments. Saints are the very exemplars of moral agents.

To state the case somewhat differently, *becoming* a saint is seen as the goal that most moral agents strive towards, but fail to perfectly attain. Insofar as we see the moral life as worthy of effort, we attempt to become saint-like, and encourage others to do the same. It seems that constraint, then, is not a necessary part of our moral judgments. Worse still for proponents of the constraint model of morality, it seems that there is some case to be made that it is preferring to act morally that is our ultimate moral goal. How could we have gone so wrong?

We propose (but certainly do no more than propose) an alternative account with which to explain our intuition that morality involves constraint. In the real world, it is one thing to know that you ought to adopt a set of preferences other than the one you currently have, and quite another to actually internalize these preferences. Sara knows that she ought to prefer to eat more vegetables, but darn it, veal is so much more appetizing! In the real world, in order to come to prefer something different from what we do currently, we must modify our behavior somewhat. We might hang out with people who currently love what we do not, in an effort to find something desirable in it. We might avoid that which we currently like, in an effort to overcome our attachment to it. We might have someone attach electrodes to our body, in an effort to negatively reinforce behaviors detrimental to this change we wish to internalize. In short, we modify our behavior patterns in order to modify our preferences. In moral cases

(ignoring whether or not Sara's veal preference is a moral issue) this changing of behavior would look suspiciously like constraining the maximizing pursuit of the satisfaction of our preferences. But a straightforward maximizing account may be made for this apparently constrained (irrational) behavior. Upon coming to the conclusion that it is worthwhile (in our interest) to change our current preferences, we modify our current behaviors in order to do so. And so we see no need for the constraint of self-interest to be a necessary part of acting morally.

The case seems at least as appealing in the case of our encouraging others to be moral agents. Jan Narveson has pointed out that one of the features of a moral system is that agents who engage in actions that are required by it are praised, and agents that engage in actions that are proscribed by it are blamed for so acting. This is a part of the informal, uncentralized reinforcement of the dictates of morality.[164] Such reinforcement aims at changing the behavior of immoral agents, and reinforcing the behavior of morally inclined agents. In short, decentralized rewards and sanctions seems to be used in part to make it the case that agents who are not saints eventually prefer to act morally.

At first blush, one might suggest that people engaged in such activities are encouraging people to constrain their behaviors. But the use of praise and blame, and more generally reward and punishment, is better understood as an exercise in behavior modification. Behavior modification differs from mere constraint because of the results over time. By merely constraining a river from running down some section of its bed by damming it up, I do not thereby expect that, upon removing the impediment the river will thereafter fail to rush through that area. But, by digging an alternative bed that is deeper

---

[164] See Jan Narveson, *The Libertarian Idea*, Temple University Press, 1988, esp. p. 125. Also see section 3.22 of his essay "Remarks on the Foundations of Morals", forthcoming.

and wider than the original riverbed, and channeling the river through that bed, I thereby expect that the river will run through the modified route. Moral (re)education aims to change the preferences of people through behavior modification. The hope is that such praise and blame will no longer be necessary, that eventually the agents will prefer to act morally.

Having argued for the plausibility of an orthodox rational choice theoretic account of morality, it may now be appropriate to briefly detail some of the largest problems endemic to a dispositional account of morality. In such a way we hope to present not only a defense of the plausibility of an orthodox rational choice contractarian account that ignores the need to explain constraint's place in moral theory, but also show that this type of project avoids certain difficulties thought fatal to the type of proposal favored by Gauthier.

In the first place, there is the problem of how often one ought to engage in deliberation regarding the rationality of one's disposition. If it is an effort that is undertaken only once in one's life, then it is difficult to understand how one could possibly hope to rationally calculate the correct disposition to adopt.[165] Recalling Danielson's recognition that the disposition you should adopt depends on the dispositions adopted by the rest of the population with whom you expect to interact, and that the population is likely to change over time, any calculation made over so long a period of time seems rationally suspect. Moreover, there is the question of when it would be rational to engage in this deliberation, assuming that agents have a choice about this. Given the foolishness of moving first in any game where predation is possible, choosing

[165] This general objection lurks behind many of the objections MacIntosh makes to Gauthier's dispositional account of rationality.

a disposition before anyone else in the population does also seems ill advised. No solution to this puzzle is likely to be forthcoming. Requiring choosing a rational disposition any other number of times is likely to be either an *ad hoc* stipulation, or to depend on the claim that the correct number of times to choose a rational disposition is the rational number of times – that is to say, the number of times that it would be utility-maximizing to do so. Such a project quickly becomes much too complicated to compete with the simpler alternative we endorse above. This is so especially when one considers that if the frequency of deliberation regarding a rational disposition becomes too great, the compliance problem re-emerges. If it is rational to re-assess one's disposition too regularly, one could be a constrained maximizer at the time of an agreement but re-emerge as a straightforward maximizer in order to defect when it comes time to act.

We add to this MacIntosh's observation that if a disposition is understood as being activated in certain situations, and forcing an agent to choose differently from how they otherwise would (in such a way as to maximize their expected utility directly) then dispositions act as an *involuntary* constraint. The text quoted above does support such an interpretation. Insofar as this concerns the adoption of the disposition, this seems unproblematic: the adoption of the disposition is voluntary, and so may be considered moral. But this is not constrained behavior, and so on Gauthier's own terms would fail to be genuinely moral. In situations where the disposition is engaged, however, this would entail that the behavior exhibited by the agent was not voluntary, though it would be constrained behavior. Insofar as we think that an action must be voluntary in order to be moral, dispositions might be thought to ensure that any action resulting from constraint was not moral behavior. There is no such problem with the orthodox contractarianism

outlined above. In short we conclude that together these problems are sufficiently significant and complex, to suggest that the simpler alternative ought to be embraced.

III: Examining Gauthier's Project

Now that I have both defended the orthodox account over Gauthier's proposal, and shown how one may translate the results of the latter into something useful for the former, the time has come to see what Gauthier's efforts have to offer to our central problem. From the revisions to Gauthier's account, we have concluded that those interested in maximizing the advantage to be derived from co-operative encounters, in which they will either be pitted against straightforward or constrained maximizers, would rationally choose to adopt preferences regarding how to interact in these situations. More specifically, they adopt preferences to fairly co-operate with others who also so prefer, and otherwise to not co-operate when others would prefer to defect on the agreement. As Danielson points out, one significant weakness of Gauthier's program is that it does not consider all possible patterns of interactive preferences.[166] Insofar as Danielson's and Gauthier's projects are considered comparable, Danielson may be understood as attempting to correct that deficiency, although he, too, obviously falls short.[167]

We do not think this so serious a charge as Danielson does, although we suppose that this is so for particularly biased reasons. Our contractarian proposal made in chapter two makes it quite clear that what we are interested in obtaining is some method through which to provide a rationally compelling justification of a morality to an existing population. Not all possible motivations are of interest to us, only the motivations of the

---

[166] Danielson, *Artificial Morality*, esp. pp. 13-14.

[167] As he recognizes at the end of his book, given his open invitation for others to continue to test various models. See Ibid. p. 202.

existing population. More correctly, it is the production of a compelling morality while avoiding reliance upon an exhaustive account of these motivations that here concerns us. In any case, we are not concerned with the extension of this problem to all possible populations.

More significant is the conditional nature of Gauthier's argument. Gauthier's argument captures only those people interested in strategic interactions, and in maximizing their expected utility in some of these encounters. Not all people can be expected to share this interest. Some may prefer self-sufficiency, and we may not dismiss the stoic lifestyle out of hand. Gauthier will not find this problematic, as morality as envisioned in *Morals by Agreement* covers only market activities, aiming to correct only market failures. For Gauthier, merely occupying the same environment and making use of the same resources is not yet enough to suggest that morality has a place. We define an interaction more broadly than Gauthier's market interactions and claim that morality may have something to say about these as well. An interaction is one in which some person's utility is affected by an action of some other person. As suggested in chapter 2, we cannot yet rule out the possibility that moral theory has something to say about all such interactions. Since this is so we cannot discount the hermit from possible inclusion in the moral sphere.

Hermits and Stoics, we may reasonably suppose, are not trying to maximize utility in market transactions, and so cannot rationally be convinced by any proof dependent upon the assumption that they are motivated to do so. This is not to say that Gauthier's account holds no appeal. Insofar as someone is interested in maximizing the benefits of actual, imperfect, market transactions, Gauthier provides good reason for them

to adopt the proposed preferences. There is further good reason to suppose that this interest is almost paramount; market societies abound and those failing to be market societies still evidence significant co-operative behavior between individuals. Common phrases like "everyone can use a friend", "no one is an Island" and "don't burn your bridges" all attest to the importance that people place upon the co-operative surplus which may be generated by working well with others. Gauthier's strength is Hobbes' weakness: Gauthier recognizes as significant the interest that most people have and that is often effective in action, namely, the interest in productive cooperation. Hobbes' strength is also Gauthier's weakness: Hobbes', if correct, identifies a trait that is universal amongst the population, albeit not one that is necessarily effective in action. It is to see what we may glean from the two proposals that will hereafter occupy our attention.

Chapter 6
~ Where we end up ~

I. Justifying the Contractarian's Efforts.

A common phenomenon in any course that introduces students to moral theory is the inevitable espousal of ethical relativism by one of the more outspoken students in the group. "But why are we bothering with all this stuff?" they ask, "It's all just what you feel is right anyway, isn't it?" Upon being asked what they mean by that remark, students tend to respond that the right thing to do is either: whatever you want (ethical subjectivism), whatever you feel is right (conscience theory), or whatever your society tells you is right (cultural relativism). Responding by counter-example, one can generally overcome ethical subjectivism by pointing to particular people in history who (almost) everyone would call immoral, but who were nevertheless simply doing what they wanted. The merest mention of Adolf Hitler, Joseph Stalin, or The Marquis de Sade is usually sufficient to convince people that ethical subjectivism is not a position that they can reflectively endorse. Pointing out that if it were the case that any of these people *thought* that they were doing the moral thing then conscience theory would proclaim that they *were* doing the right thing is also usually sufficient to convince people that they are not supporters of this position either. Cultural relativism is usually overcome by citing example of extremely repugnant behaviors systematically recommended by existing societies. Pointing to societies that engage in female genital mutilation or the killing of unwanted children by leaving them to die of exposure tends to encourage people to closely examine alternative moral theories.

But the lecturer who has so responded has engaged in a useful (some might even say necessary) sleight of hand. Pitting the student's moral intuitions against the cited

persons or societies, and finding that they diverge, does not show the falsity of ethical subjectivism, conscience theory, or cultural relativism. It may be that all that this shows is that each student so convinced personally disapproves of the person or culture. From such disapproval it in no way follows that any of the subjective moral theories are false, but only that the disapproving student does not consistently espouse ethical subjectivism. The perceptive student recognizes this, and one occasionally receives a response to that effect. Mere intuition divergence in no way disproves subjectivist ethical theories. This perceptive student has picked up on the general acceptance of skepticism regarding the validity of moral theories; it seems unlikely to most that one could rise above the mere clashing of intuition against intuition when discussing the moral standing of particular activities.

This skepticism has led to a shift from the identification of what the correct moral theory is, to a series of debates regarding what the right thing to do is, in particular circumstances. In political theory, efforts have shifted away from attempts to justify a particular institutional arrangement, and have focused instead upon how one ought to be treated in such arrangements. Rawls' progress may be taken as a telling example of, and perhaps a significant contributing factor to, this shift. *A Theory of Justice* attempts to justify a particular theory (justice as fairness) over another particular theory (utilitarianism). In *Political Liberalism* Rawls shifts his emphasis to see instead what institutions would maximally realize the ends of a liberal democracy. If your moral intuitions support liberal democracy, then *Political Liberalism* may have something to say regarding how to best realize your supposed moral state of affairs.

Or perhaps not. The force of these efforts rest on presenting a theory regarding how to act in certain circumstances that follows from widely and strongly held intuitions. Unfortunately, for this and other similar efforts, it seems that no such intuition exists, or does not do so in a vacuum. People seem divided regarding the most basic moral intuitions. Theorists who fiercely cling to the moral intuition that people ought to be treated with respect tend to be pitted against theorists who just as fiercely cling to the intuition that people ought to be left alone. This comes to the fore most clearly in arguments regarding redistributive taxation – one side concluding that taxation ought to be allowed to the degree to which it is necessary to ensure human dignity, and the other strenuously denying exactly that. People who seem to agree on a particular intuition disagree regarding its strength relative to others. Those who agree that a person ought to keep her promises, and that a person ought to help people out of dangerous situations, nevertheless may also disagree about whether to keep a promise to help Kathryn study for her final exam (which she is sure to fail without such help) or instead give aid to a dazed pedestrian, helping him safely through the last few blocks to his home. And finally they may disagree regarding what some commonly held intuition entails: it may be that both theorists agree that the dignity of persons must take precedence, but that the one argues that people's dignity is best guarded by allowing them to keep that which they have gathered, and that the other argues that people's dignity is best guarded by passing on what others have gathered to those who have not been able (or willing) to gather. Appeal to intuitions has largely led to intractable debate. This, in turn, has led to a deepening of the skepticism that was partially responsible for the adoption of such argumentative styles in the first place.

There are two problems in evidence above: disagreements regarding the intuitions, and disagreements regarding the implications of the intuitions. The latter is not our bailiwick. These disagreements find much of their root in disagreements over particular facts – whether dignity is best guarded by redistribution or not is clearly an empirical matter, assuming that one has clearly defined what dignity is. A justification of particular intuitions, however, is our bailiwick. Mere reliance on intuition has not furthered the moral debate, for intuitions are insufficiently shared. Moreover, even if such intuitions were sufficiently shared, such a method of justification still allows for the possibility that our answers might be systematically mistaken. It is only egotistic posturing that allows us to pretend that we are immune to the kinds of mistakes with which we have charged the slave-owners, baby killers, and woman mutilators.

So what is our alternative? We must attempt to rise above the mere clashing of intuition against intuition however unlikely the success of such an enterprise is supposed to be. We must provide a method with which to justify some set of moral intuitions in the face of all contenders. Pessimism over the possibility of justifying a particular moral theory drove theorists towards a more limited model in which to argue in order to provide what answers may be gleaned from reliance upon commonly held intuitions. Such attempts at justification have become the norm, and efforts to justify particular systems have largely fallen by the way. The recognition that the moral debate which relies on intuitions has become intractable gives a new urgency to the justificatory enterprise. No good advice can be given that ends with the phrase "...or at least, that's what I'd do!" Moral advice that relies for its force upon the holding of a particular intuition is not really good advice at all. This advice is no more than an opinion, and to rely only on opinions

really is to suppose that Hitler's actions (assuming that he did not privately think that he was doing the wrong thing) are on the same moral level as any other person's. This supposition renders the enterprise of moral theory pointless.

Have we here merely engaged in another sleight of hand? Not at all. Finding it undesirable to allow that reliance upon moral intuition is all that there is to the moral debate, due both to the intractable (and thus useless) nature of that enterprise and the repugnant conclusion that a sincere Stalin was a moral person, we therefore attempt to justify some intuitions over others. We attempt to provide a justification that may, in effect, raise the status of some intuitions to the level of moral dictates. That is not to say that this effort aims to justify a particular set of moral intuitions. It is clear that to attempt a biased justification is to end up providing no justification at all. Our aim is to overcome the skepticism regarding this justificatory enterprise so that the moral debate may usefully continue. It may be the case that the set of justified moral rules so identified includes moral dictates that were not widely held as moral intuitions; we will have to wait and see.

Whether the justified moral dictates were widely held as intuitions or not, whenever a person's moral intuitions diverge from the set of rules identified by our efforts, the justification must provide sufficient reason to set those intuitions aside, and embrace the justified morality. Only in such a way will this justification overcome the problem of intractable debates described above. It is only through a recognition of the superior nature of the justified morality over any particular person's intuitions regarding how to act that one can expect any particular person to be guided into so acting. And morality is about advising one how to properly behave.

And so a fundamental justification must identify a compelling set of moral dictates in a non-biased way, if it is to have any chance at convincing people that this identification is also a justification. This requirement is forced upon us given our project; if we are attempting to further the moral debate, to provide reasons for certain actions or policies, we must provide convincing arguments. The justification must, in other words, give some reason for some theory's adoption, if it is to be a justification, as opposed to merely identification. It must produce non-question-begging reasons for the adoption of the set of dictates that it identifies as those of morality. Given our pre-theoretical understanding that moral theory has to do with recommending certain actions, a fundamental justification of moral theory must have practical force. Our fundamental justification must provide reasons for acting in accordance with the identified dictates of morality. These requirements, derived as they are from the nature of the enterprise, are in turn also dependent upon a certain view of morality as social construction, as opposed to a rational intuitionist conception of morality, whose likely end is the skepticism we are trying so desperately to avoid.

John Rawls describes rational intuitionism as the view that is characterized by four general features: true moral first principles correctly describe "an independent order of moral values", these principles are known through theoretical (as opposed to instrumental) reason, they rely on the mere recognition of the principles as sufficient to motivate one to act in accordance with them, and a statement is true when it accurately describes the object being discussed.[168] This is certainly an adequate representation of the view that contractarians must reject. Morality is not seen as an independent entity existing in metaphysical space, waiting to be discovered and governing action thereafter.

Morality, it is supposed, is both part artifice and part artifact. Human interaction, both intentional and otherwise, has proceeded in such a way as to give rise to morality.

Given that we suppose that human interaction has given rise to morality, we may turn to our primary motivation in order to both discover from whence morality resulted, and from whence morality is to be justified. Given that we are driven by our preferences to purposeful action, and that morality is supposed to be a result of human actions, we may conclude that moral practice arose, in large part, due to our preferences. Supposing that our preferences primarily govern our actions also has the effect of allowing our justification of morality to proceed from the same domain from which the practice arose. The need for a justification arises because our actions are also driven in part by affect, and our reasoning skills are imperfect. These imperfections obscure the logic of morality.[169]

Without this obscuring of the end of morality, it would be perfectly obvious what reason we have to adopt moral behaviors. With the observation that people are imperfectly rational and have affected behaviors, we may also expect that particular moralities that exist in society may not maximally achieve their goal; they may fail to maximize the satisfaction of people's preferences given interactions with others in various real-world circumstances. This allows the theorist to expect some separation between actual practiced moralities, and the identification of a critical moral theory that is maximally compelling – a theory that would, if presented to people, actually be thought to have the best chance to maximize preference satisfaction. In such a way we can speak

---

[168] See esp. Chapter 3 of *Political Liberalism*.
[169] Gauthier makes a similar point in *Morals by Agreement*, p. 60-61.

of a justification of morality and hope for a correction of some positive (existing) moral codes.

And so we have come to conclude that a fundamental justification of moral theory must rely upon people's preferences; that in order to present compelling reasons for the adoption of the tenets of morality, we must rely upon peoples' preferences. In order for this justification to be maximally compelling, all preferences must be included. Preferences will both identify and justify morality. Insofar as people are rational agents, we can also be hopeful that the theory so identified will be reasonably close to our common sense understanding of the content of the tenets of morality. Insofar as any divergence will be due to people's irrational behaviors, we all have reason to prefer the theoretically derived moral theory to any historical artifact.

II. What form the Project must Take

The contractarian, then, is interested in identifying a social choice (the rules thereafter governing behavior) via individual rational choice (which set of rules each person prefers to so govern). Social choice via individual choice, given cardinal utility, is derivable either from amalgamative models or bargaining models. People's group choices over action are either the result of adding up each person's preferences and finding out what would give maximal satisfaction, or they are the result of bargains struck - each person coming to agree with the others over the actions to be performed by each person. In the real world each procedure is commonplace; political decisions are by and large made via some calculation over what would produce the greatest good, and market transactions (analytically) involve bargaining scenarios.

The case of the identification of moral rules, however, is a different animal entirely from the exchange of some amount of money for a saucepan. In the case of the exchange of money for a saucepan, each agent (the shopkeeper and the would-be-cook) is able to have a rough and ready understanding of the future benefit of engaging in such a transaction. The cook has some idea of how much better off he expects to be given that he would have a new saucepan at his disposal, and be out some amount of money. The shopkeeper has some idea of how much better off he would be with some more money at his disposal, but be out one saucepan. These expectations depend on certain background conditions likely remaining the case. The cook who already has a perfectly good saucepan at home is likely willing to give up less money than if he had no such item. The shopkeeper would be less willing to exchange the saucepan for some amount of money if it was the case that the addition of this amount of money would put him into a new taxbracket, and therefore result in his having less take-home cash. It is the expected stability of these background conditions that makes it reasonable to ascribe a particular expected utility to any given trade.

In the contractarian endeavor, however, none of the otherwise background conditions could reasonably be assumed to remain the same. For each piece of property you have at your disposal, there is a chance that it will not remain at your disposal. Non-taxation is not at all guaranteed, to say nothing about the particular form of taxation that would be instigated. It may also be the case that practicing philosophy will be banned, in which case students of philosophy will be worth less to others than they might otherwise be. In short, no one can reasonably expect to come to any reasonable conclusion regarding the relative worth of each participant's proposed contributions. As a result of

this, no rational bargain can be struck. It would be extremely unreasonable to suppose that an agent would engage in bargaining activities given no reasonable expectations of the relative value of each other person's contributions, and the relative cost of their own proposed contributions. The only possible reason for such a bargain would be that there was no other way to come to a decision regarding the social choice of the tenets of morality, and that coming to any decision at all was thought by all to be in their interests, all things considered. As mentioned above, however, this is not the case. The amalgamative option is an available option, and it is to this option which we now proceed.

Amalgamative contractarianism seizes on the motivational force of all of each person's preferences and beliefs about any given set of rules' effects on the world, and with these in hand ranks each complete set of rules governing interaction in relation to each other. These results are tallied together, and the maximally preferred system of rules is identified. This should not be understood as a second-best option. It is not only endorsed because of the bargaining model's failure to produce useful results. It is not obviously the case that most people in any given population would prefer the bargaining model to the amalgamative model. There may be good reasons to suspect that, *prima facie*, those who would have greater influence should a bargaining model be made use of instead of an amalgamative model would, other things being equal, prefer that such a model be made use of if it could be. But this has not yet taken into account whether all other things are equal. If any given population has independent positive preferences regarding, for example, the democratic commitment to "One person, one vote!", then these *prima facie* reasons would not necessarily be found decisive. Even if the

bargaining model is possible – which we have argued is unlikely – it is perhaps even then not always ultimately compelling.

There is one glaring failure common to all models which attempt to provide an ultimately compelling justification for moral theory. Finding that the inclusion of a greater number of people's preferences in a justification tends to increase the reasons that that person has for finding the justification sufficiently compelling, and concluding that this implies that the ultimately compelling justification will thus be one that includes all people's preferences, we find the proposal impractical. While the calculations required of this proposal are not too computationally complex to be undertaken, the inputs necessary are lacking. We do not have at our disposal an exhaustive list of each person's preferences and relevant beliefs about the world. No exhaustive list is likely to be forthcoming either. Instead of admitting defeat, or at least a temporary withdrawal from the field, we proposed to examine what characteristics a partial list of preferences must have, in order to provide a sufficiently compelling justification for us to overcome the reliance on mere intuition in moral debate.

III. What we have learned

III.i Thomas Hobbes

Thomas Hobbes presents us with a deductive account of what drives people to action. Starting with the principle that an object in motion stays in motion and the supposition that living consists of the moving of parts of a body instigated internally, Hobbes eventually concludes (among other things) that people continually tend to strive for what they find useful. It is argued that the recognition of this striving, and the

scarcity of things useful, will eventually drive people to conclude that it is reasonable to expect to come into conflict with each other. Finding this unpalatable, it is then argued that it is reasonable to instigate an all powerful Sovereign who will change the situation sufficiently that it is reasonable for people to foresee no such conflict. We did not examine Hobbes' argument for the Sovereign, as it is not his conclusions that here concern us. We are interested in his deduction of a preference that is always at work in all agents.

We may attribute to Hobbes a minimal definition of an agent: an agent is that which has the power to move its parts through processes internal to it. Examining that type of agent most likely to correspond, in the relevant respects, with people, he presents a mechanistic process that originates in the senses.[170] The motions of the world react with our bodies, via the senses, in such a way that fancies (immediate perceptions) are produced in the mind. These fancies, being motions, continue to exist in the mind long after the initial sensory experience, and are called imaginations. Each fancy produces an endeavor in our mind. These endeavors are motions either towards or away from the cause of the imagination (either appetites or aversions.) These fancies are the beginning of voluntary motion.

Hobbes also details our thought process. A train of thought is a succession of imaginations. The past succession of fancies guide our manipulation of thought, each segment in a train of thought having been joined with its predecessor and successor previously through the senses. These trains of thought are of two kinds: regulated and unregulated. Unregulated thought does not here concern us. Regulated thought takes

[170] We ignore for the sake of simplicity the problem of the status of Hobbes' proposed natural appetites, which he identifies on pages 119-120 of *Leviathan*. We remarked on this difficulty in 4.I.ii.

only two forms: that of seeking and that of inventing. Seeking involves imagining possible ways to bring about a particular effect. Inventing involves imagining all of the possible ways in which a particular object may be used. Neither of these types of regulated thought are ever idle speculation. They always aim at some end identified by a passion - the obtaining of the object of an appetite, or the avoiding of the object of some aversion.

Hobbes concludes from this description that, as long as an agent of such a constitution exists, it will tend to continually acquire power after power. A power, it will be recalled, is that which is thought useful in acquiring a future good. Given that fancies will always be produced, and that regulated thought will, in part, tend towards discovering ways to acquire that for which one has an appetite, it may with some justification be concluded that these agents do have this tendency.

But as Hobbes himself makes clear, each regulated thought does not, by itself, lead to action. It is only the last appetite or aversion in any deliberation that is identified as the will of an agent - as that which leads to action.[171] It is, ultimately, courses of action for which we wish to successfully lobby. What we cannot rely upon is a mere inclination. Inclinations may be overcome. There is nothing to require that a tendency ever once be effective in action. In searching for an ultimately compelling justification, we must search for preferences that are of paramount import. It is only in such a way that we can maximize the justification's efficacy.

---

[171] See esp. p. 127 of *Leviathan*, where he states "...the whole summe of Desires, Aversions, Hopes and Fears Continued till the thing be either done, or thought impossible, is that we call DELIBERATION. ... In *Deliberation*, the last Appetite, or Aversion, immediately adhæring to the action ... is that wee call the WILL".

III.ii David Gauthier

David Gauthier presents us with compelling reasons to accept that agents, who find it to be in their interest to engage in market behavior and find themselves in a perfectly competitive market, have no need of the council of moral theory. In a perfectly competitive market each agent is affected only voluntarily, and voluntary exchange and consumption tend towards equilibrium. Each person will end up satisfied with their bundle of goods over any other possible bundle of goods which could be received though trade, given the rate of exchange. It is also the case that this equilibrium state is optimal; no one could be made better off without someone being made worse off. Ignoring for the moment any claims regarding the fairness of the initial distribution of goods – claims which Gauthier has made some argument regarding – what need would such people have of moral counsel?

In situations where co-operative outcomes provide some benefit, and some benefit is also to be had from not acting in compliance with the proposed co-operative outcome, equilibrium and optimality do not coincide. When the possibility exists to trade some amount of money for a saucepan, assuming the participants are interested, there is some benefit for each if an agreement can be reached regarding the amount of money to be exchanged (say, $50.) Yet there is still (other things being equal) some advantage for the shopkeeper to attempt to retain his saucepan, and grab the $50. There is also some advantage for the cook to attempt to retain the $50, and grab the saucepan. In such situations, let us suppose, the result is that neither brings their item into reach of the other, and as a result no exchange takes place. For either agent to act differently would be ill advised, as bringing the item in question into the reach of the other allows him or

her to attempt to take the offered item without engaging in the exchange. And so equilibrium and optimality come apart. Each would prefer that the exchange take place, and would be better off if it occurred, but by each acting in such a way as to maximize their expected utility, no such exchange takes place; by each acting in a utility maximizing manner, both fail to end up in an optimal state.

In coming to recognize the existence of these situations, our reinterpretation of Gauthier's arguments suggests that agents have a choice to consider regarding the preferences that they choose to adopt. They can either allow their preferences to remain the same, or they can adopt preferences in which they prefer to act on co-operative arrangements when they think that their possible partner in the interaction also so prefers, and otherwise not act on these possible arrangements. If they choose to keep their original preferences, then they will not be able to engage in exchange. Straightforward maximizers will either interact with other straightforward maximizers, or (the now unfortunately named) constrained maximizers. If interacting with other straightforward maximizers, no trade will take place, for the reasons outlined above. If interacting with constrained maximizers, no trade will take place if the constrained maximizer correctly identifies with whom he is dealing. Constrained maximizers, however, when interacting with other constrained maximizers, will prefer to co-operate, and so the exchange will take place. Optimality and equilibrium are rejoined in these situations, and so we conclude that these arguably moral agents are all better off, and morality has fulfilled its function.

But this argument rests crucially on the assumption that only agents wishing to engage with each other in market interactions are the agents to whom morality is to

apply. We take issue with this assumption. It is definitive of this enterprise that it attempts to cast the net of morality as widely as possible, and Gauthier's approach seems a hair too thin. Agents who do not so wish to engage others may still be engaged in interactions with others. We defined an interaction as one in which some agent's utility is affected by some other person's action, and that each person engaged in an interaction is either affecting some other person's utility, or being so affected themselves. Morality, it was supposed, exists to regulate such interactions – either by permitting, requiring or forbidding them. Given the nature of this project, Gauthier's proposal does not capture each agent to which morality must be justified. And so we insist that the preferences identified as sufficient for our project must not only be of paramount import, but also be of paramount import for each agent to which morality applies.

III.iii The Possibility of Paramount Preferences with Universal Scope.

When suggesting that the preferences to be made use of must be of paramount import, and that we must find preferences of paramount import for each agent, one might naturally conclude that what we need to discover is one preference that is of paramount import for each agent. While the discovery of such a preference, if made, would indeed be sufficient to provide the hope of a successful conclusion to our enterprise, we hold no hope that it will be made. Given that preferences, given beliefs, are what lead to action, should such a preference exist, we expect that most choice made in the world would provide evidence for its existence. If it were the case that every person on earth had an incredibly strong preference that they stay married, we would expect to see that divorce almost never occurred. If it were the case that every person on earth desired their

personal liberty over everything else, gender oppression and submissive sexual games would be almost unheard of. If all people were driven by a respect for persons, holy wars would not be waged.

There is little as obvious about people as the fact that they differ in their preferences. When we suggest that preferences must be found for each agent that are of paramount import, we are not suggesting that hegemonic preferences, universal in scope, need be discovered. We are instead simply requiring that hegemonic preferences be identified for each person to be guided in interaction. We also need not be committed to providing only one preference for each person to be so guided. What the contractarian needs is a list of preferences for each agent strong enough that however incomplete this set is, each agent will be sufficiently satisfied by the maximization of the consideration of this set of preferences in justifying the dictates of morality so as to overlook that the list made use of was incomplete. It is supposed that it will so satisfy people because those preferences that were not included were, by definition, not very strong. If they were not strong we can conclude it likely they would have made little difference in the substance of the rules so identified. There is nothing in this aim that requires that only a single preference be made use of for each agent.

This is fortunate indeed. Were it the case that only a single overriding preference had to be made use of, our enterprise would have little chance of success. If it were the case that, for most people, a single preference was decisive in action, then people would be much easier to predict than we commonly take them to be. People, however, quite often surprise us in their activities, and many of these activities are not attributed to a single preference. When explaining a choice to become employed in a particular

profession, one tends to say things like "Well, the salary was good, and the benefits were fine, and I couldn't beat the three weeks paid vacation!" People are complicated, and it is unrealistic to suppose that in most cases a single preference will be found that is so important to a given person that its inclusion in the contractarian calculation will satisfy her.

And so we are left with the task of identifying some set of preferences for each agent such that their inclusion in the contractarian calculation would produce an ultimately compelling justification of the tenets of morality. That this is our task, as opposed to simply identifying a universally paramount preference that all people hold, means that we must go beyond a mere identification, and discover both the distribution and strength of each preference. For any particular set of preferences identified as being of sufficient import for a segment of the population, we would also need to identify what percentage of the population held the inclusion of that particular preference set to be sufficiently compelling. Without such a figure, mere identification of the preference set would not be worthwhile, as it could not be meaningfully incorporated into the contractarian's efforts. Let us suppose that it has been found that some members of the population hold preference set A to be sufficient. Without an identification of the number of people who hold this set sufficiently strongly, we have no idea how much weight to give this set of preferences when engaging in the contractarian calculation. We would need to find out not only that preference set A is held, but also how many held it. Only upon discovery that 25% of the population held A, and 25% held B, and that the other 50% held C, would we be able to proceed. Without such information, we would not be able to effectively calculate the weight each preference set is to have in the

calculation. Similar reasons require the relative strength (cardinal measure) of each preference in each set, in relation each other member of the set. Without such weighting, no meaningful calculation can proceed, as was made clear in chapter 2.

Is it reasonable to be discouraged, given these requirements so set out? We think not. These requirements, while substantial, are not insurmountable. David Gauthier identified a large segment of the population when he suggested that morality applies only to agents who are interested in cooperating with each other for mutual gain. This group grows even larger if one includes those who are willing to tolerate each other (although not necessarily engage in market exchanges with them) for the peace that it provides. Adding to these the groups who wish to remain isolated from others (the stoics mentioned in 5.III) and those who wish to wage war on those different from themselves (for example, for religious reasons) and the net seems to capture most imaginable people. This is offered, not as a considered proposal regarding how to capture everyone's interests, but instead as an example meant to suggest that the chances of success are not so bleak as might otherwise be assumed given the conditions we must meet as outlined above. It is not solely through conceptual analysis that the preferences of people are to be described, but by the linking of conceptual analysis with empirical evidence. But through what means may we expect to find evidence to which we may apply our analytical tools? It is this matter which concerns us in our penultimate section.

IV. Where we should Look

IV.i Evolutionary Game Theory

One possible area for exploration opens up for us because of our rejection of the necessity of constraint for moral activities. In chapter 5 we argued that constraint is not a

part of moral activities, based in large part on our intuitions regarding the moral status of saint-like behavior. As outlined in chapter 1 Peter Danielson rejected some evolutionary approaches to moral questions solely on the grounds that they would never be able to produce any recommendations that suggested constrained moral behaviors. Perhaps evolutionary game theoretic constructions could offer us some responses.

Unfortunately, it is immediately clear that such an enterprise is not empirical. Such constructions could tell us that *if* agents had preference sets A, B, or C, and *if* they were interacting in games that had such-and-so characteristics, *then* the agents that had preference set B faired best. They might also suggest that agents which have preference set C were least likely to be invaded by predators, or that they would all be well advised to change their preference sets to sets A2, B2, or C2. What evolutionary game theory does not do is provide any information regarding the preference sets that people actually have.

One form of this project tries to explain how it is that we have come to have the preferences that we do. Starting with a set of preferences designed to mimic the preferences we might expect to have at a lower level on the evolutionary scale, this project attempts to show how it is rational to have developed into people who have the preferences that we have now. How it is that, given our environment and interests, we have developed into the people that we are. The search to explain the evolution of altruism is of perennial interest to evolutionary game theoreticians, and stands as a paradigm example of such projects. It is clear, however, that these projects can only succeed if they have independently arrived at conclusions regarding the preferences we might have at a lower level on the evolutionary scale, and the interests that we have now.

In short, they require what we require, a list of preferences that we actually have. It is empirical work alone that will produce such results, and it is the empirical sciences upon which we must focus our attention.

IV.ii Historical Analysis

An historical analysis of human motivation also holds some initial appeal. Perhaps we ought to examine what we have noted regarding past human efforts in an attempt to find anything consistent in it which could provide the basis for a contractarian justification. Hobbes himself has been thought to rely heavily on this method by way of assuring a close fit between his deductive efforts and the problems he saw in the real world, with which he was primarily concerned.[172] There is more than just historical precedent to recommend this method, of course. A successful survey of human motivation over time would return with data of preferences that led to actions which are stable over time, which could be taken to indicate both the universal applicability of the sets discovered (in the sense outlined above) and that these preferences are strong enough to lead to action.

We must, however, carefully examine the strength of any conclusions based on historical analysis. Any pattern of behavior identified is likely to be only a general pattern of behavior, indicative perhaps of only the 'noteworthy' of humanity. Any historically-based pattern of behavior amounts to no more than a pattern of behavior to be found in those described persons. History is radically incomplete, and the results of combing it for a general motivation for behavior must, of necessity, be incomplete as

[172] As he states in his conclusion "And thus I have brought to an end my Discourse of Civill and Ecclesiasticall Government, occasioned by the disorders of the present time..." He has been argued to turn

well. This is not to say that the pattern of behavior so identified would be found not to hold for people in general, but only that other means must be made use of to see whether this is so.

History is also an inexact discipline. The ascription of conflicting motivations to particular agents who have engaged in particular actions is commonplace, and problematic in the extreme. If we cannot turn to history to give us clear answers regarding the preferences of people, then it may not be able to help us. It may be said in reply that history can help to identify certain patterns of behavior, and that its strength lies in providing a large-scale analysis of human interaction. To expect exactness at the individual level is perhaps not only unreasonable, given the natural limitations of the subject, but also unnecessary. If history can identify general patterns of behavior, then we can make use of it to provide contractarian justifications for particular moral dictates.

Ignoring for the moment that the substantial problem that ascribing general patterns of motivation is also faced with a problem of multiple possible descriptions, historically identifying general patterns of behavior is still not sufficient for our contractarian efforts. A pattern of behavior so identified amounts to an identification of a tendency in people to act in such-and-such a way over long periods of time. But a justification must deal with particulars. It is to be justified to person X, person Y and person Z, who at this moment may not act so as to conform to this pattern. People at particular times are different than at other times, and patterns of behavior are necessarily abstracted from the particulars. If a person endorsed pacifism over 65% of his life, and thought seriously of being violent only when first attacked over 30% of his life, and

---

to history to inform his efforts to examine the basic motivations of mankind, *pace* Leo Strauss, by David Johnston in the first chapter of *The Rhetoric of Leviathan*, Princeton University press, 1989.

believed in indiscriminant violence over the (presumably last) 5% of his life, an historical account would find him to be a pacifist. Any results derived from this observation would likely fail to convince him of the merits of any case presented when he was not a pacifist.

It could be responded that the above argument misses the point – any contractarian justification derived from general patterns of behavior will end up with a *likely* justification. If we can see that, historically, people tend to prefer X, Y, and Z, then any argument that relies upon these facts will end up producing a result compelling to most of the existing population. If, 80% of the time 75% of people are governed by preferences X, Y, and Z, then any particular contractarian justification that is provided by such methods would likely be quite compelling. This is not, however, a response relevant to our project.

This response would mean a great deal if we were examining the stability of any proposed contractarian justification. If we could provide historical evidence to suggest that people have been motivated by exactly the preferences of which we made use when constructing our justification, then this would suggest that the justification would remain, by and large, compelling over time. We are not, however, primarily concerned with providing proofs that our justification would result in a stable conclusion, for reasons that will be elucidated below. We are concerned with providing a justification that is ultimately compelling to existing persons. Insofar as we rely upon historical information, we are not relying on occurrent information, and this diminishes the strength of the justification. In attempting to convince actual people that they have reason to adopt some set of moral dictates, we ought to find reasons that actual people find compelling.

Lastly, we observe that people's preferences have changed over time. Acceptable behavior means different things now than it has meant before. The western acceptance of the rough equality of people's worth today stands in sharp contrast with the feudal orders of several hundred years ago. Women's rights have come a long way over the last century in some quarters of the world. Religious tolerance is also on the rise in some quarters, while some areas are backsliding into a more extreme religious intolerance. In short, people change, and reliance upon general patterns of behavior may be ill-advised given this observation. This is not to say that this observation, nor any of the previous observations, is decisive for dismissing an appeal to history. It is the combined force of all of these arguments that is decisive. Given history's incompleteness, inexactness, and questionable applicability to our enterprise, we feel that other approaches, more exacting and directly applicable to contemporary people, would be more fruitfully made use of.

IV.iii Psychological & Economic Analyses

And so we turn to two fields of investigation into contemporary human motivation: psychology and economics. We must first delineate the disciplines being examined. When we here discuss economics, we are not discussing the theoretical developments in the field of probability. Psychology is, in its turn, not to be thought to examine group market behavior in unguided circumstances and thereby derive hypotheses regarding human motivations. While economics is rightly understood to include probability theory, and psychology may be properly concerned with the examination of market behavior, we are not concerned with these aspects of the fields. We will take economics to be an examination of natural market behaviors, contrasted

with the political economics of Wicksteed discussed in chapter 2. Psychology will be understood as the examination of people's behaviors in controlled circumstances; these circumstances may include verbal responses to direct questions, but are certainly not limited to such responses. It is these domains of which we may make use.

It is not our intention to here argue for any particular form that our investigation ought to take. As we have mentioned before, it is not for the philosopher to compete with the economist, or the psychologist. It is only our intention to suggest why these domains are found acceptable while history has been found wanting. The results of economics and psychology are, of course, incomplete. Indeed, were all motivations of contemporary agents identified, the contractarian could forthwith produce an exhaustive calculation, and the discussions of chapters three through six of this project would have been entirely unnecessary. While this incompleteness is a necessary feature of history, however, it is not a necessary feature of psychology or economics. This is not to suggest that we think it reasonable to foresee a complete description of motivation for each person from either economics or psychology any time soon. Certainly we are not even optimistic about the completion of either enterprise in the near future. The point being made is one of degree of completeness. We have described history as being necessarily, and radically, incomplete. We may never know why Alexander the Great visited Siwah, nor what the priests there revealed to him, or how those revelations guided his future behavior. However there is still some hope that we will gain insight into how much pressure people feel in PC environments, and how this affects their actions in such environments. Neither psychology nor economics are necessarily incomplete, nor need they fail to capture most of what they profess.

Both psychology and economics also have some claim to rigor that is lacking in history. This again is not to assert that psychological or economic explanations are incapable of error. Both disciplines are clearly works in progress, and evolving in response to glitches made clear through experiment or repeated observation. They are, however imperfect, still more rigorous than history. And this hope of further development makes either discipline superior to an historical examination. It is through these more rigorous processes that we may eventually hope to discover what primary motivations we actually have.

Lastly, economics and psychology both more directly pertain to existing people. History, by definition, applies most directly to the past. Economics and psychology more immediately focus upon the present. One cannot run a double-blind experiment except in the present. It is true that neither discipline can entirely disregard history. A group of conclusions regarding what motivates the people of Northern Ontario will likely take as data several sample behaviors gathered over time. Psychological data is also cumulative; building on what has been concluded in the past, we construct experiments designed to further this knowledge. In neither case, however, need these disciplines focus on the past. Using past experiences, either discipline can attempt to provide us with information about the present; it is this information of which we have need.

V. What Has Been Ignored.

V.i Stability

As has been occasionally alluded to in this project, we have not been concerned with the stability of a contractarian justification over time. This is not to say that we do

not find the issue of stability concerning. We are in full agreement with those who suggest that the stability of a particular proposal is an important issue to address. We do not, however, think that neglecting this issue is a grave failing of our project. The nature of our enterprise required setting aside this issue. Any satisfactory argument regarding how to produce a resulting moral code that would be maximally stable over time would end up obscuring the logic of our arguments for a maximally compelling justification. A justification of a moral theory that makes good use of person A's set P of preferences is going to be less compelling than one that makes good use of person A's set (P + N) of preferences.[173] Arguments that depend on this line of reasoning will tend to view a person as having a set of preferences *at a particular time*, and will not consider the preferences that this same person may have in the future, or has had in the past. Justifications are presented to a person at time T_n, and the preferences that are going to be made use of when constructing a justification at time T_n, are exactly those that the person has at time T_n. This can include some preference that her preferences that she is likely to have at T_{n+1} are be given some consideration, but certainly need not do so.

Justifications that rely on stability, however, do not rely on such a conception of each person. Given a set of preferences P, it is not the case that relying on set (P + N) will generally result in a more stable set of moral dictates. It may even be the case that relying on set (P-M) would produce a more stable justification than relying on set P. Imagine the case where included in set M were very strong preferences that an agent had that striking a teenager was completely acceptable, due to the fact that at time T_{n-1} the sampled agent had just had a particularly unsatisfying row with her child on the issue of

[173] Ignoring for the moment the unusual cases where the person would prefer to not have their N preference included in the calculation.

showing basic respect for others. Such desires, while no doubt strong at the time, are famously fleeting (to the benefit of teenagers everywhere.) The agent sampled may, at most times, strongly prefer that teenager-abuse was not acceptable. Excluding set M would then produce a justification that is more stable over time than including set M. It would not, however, result in a more compelling justification at time T_n.

Those interested in providing stable justifications are going to be primarily interested in finding those preferences most consistently held over some specified amount of time. This can be specified as over the course of a person's adult life, or over peoples' entire lives, or even go so far as attempt to find preferences held over all recorded history. They need not attempt to find among those stable preferences the most strongly held, although of course they might. Compelling justifications, alternately, must uncover the most strongly held preferences that an agent has at a particular time. The aims of the two projects necessarily conflict. This is not to suggest that a complete contractarian proposal will be the result of either effort, but to suggest that the two projects should be attempted separately, and then we should see in what way they could be joined.

V.ii Limiting the Examined Population

Finding it impossible to continue with a justification that completely includes peoples' preferences, we decided to examine what would need to be accomplished by proceeding with an account that relied upon less than the complete inclusion of peoples' preferences. This strategy was not our only recourse. We could have attempted instead to produce a complete account of a smaller population. There is good reason to expect that the examination of a smaller population would more easily produce a complete list of preferences. Not least among these reasons is the sheer size of the project. The fewer

people one needs to examine, the less time it will take. Whether geographical location or conceptual criteria, (for example, all agents that prefer to engage in market activities) delineate the smaller population, one might also expect to encounter less deviation of preference from person to person.

Whether or not this last expectation could be born out, it is questionable whether or not it would result in any less effort spent on compiling the preferences of a given population. It is arguable that it would take just as much energy to identify and record two sets of preferences that happened to be identical in content as it would to identify and record two sets that diverge. This aside, we are skeptical regarding the prospects for success of any project that attempts to exhaustively enumerate any number of peoples' preferences, given the current state of the fields to be relied upon (economics and psychology.) Given that this is so, our results will be as applicable to these more specific populations. In fact, we have made no effort to exhaustively describe to whom, exactly, morality is to apply. This is not to say that we have not implicitly endorsed some conditions that these agents must meet in order to fall under the scope of morality. They must be agents, be capable of deliberative activity and of rational planning. They must also be capable of changing their set of preferences. But these conditions are not necessarily considered sufficient. If we have not come to a precise understanding of who is to be included in the set of moral agents, then it would be premature to attempt to define a further sub-set of this population.

VI. Conclusion

Finding the current stalling of ethical debate regarding practical matters unacceptable, we deny the usefulness of relying on moral intuitions in debates about the

proper course of action. The only alternative is to rise above this useless yammering, and renew our efforts to justify a particular set of moral dictates. As a justification must provide reasons for the effective adoption of a particular moral code over its contenders, we must base our justifactory efforts in that which motivates people to action. Finding economic concerns to be significant motivators, we base our justification on economic concerns. We have some reason to suppose that the result will be a recognizably moral theory because of our assumption that our common understanding of morality arose through human interaction – insofar as humans are rational, the results of human interaction are supposed to be so as well.

Danielson and others have convincingly argued that traditional rational choice theory is incapable of providing a justification of morality that would be a recognized constraint on otherwise straightforward maximizing behavior. We must therefore either reject traditional rational choice theory, or reject the intuition that morality must involve constraint. Finding the intuition insufficiently grounded, and providing some explanation regarding how it may have been mistakenly acquired, we choose to reject it, and embrace instead traditional rational choice theory in our efforts to provide a rational basis on which to justify a moral theory.

Justifying a particular moral theory is effectively also endorsing a particular pattern of behaviors as desirable. Basing such an endorsement on individual preferences is attempting to derive social choice from individual choice. We examined two different ways in which to attempt such a derivation: bargaining theoretic or amalgamative efforts. Bargaining theory cannot produce a rational account of an agent's expected worth to others. Any attempts to stipulate expected worth deviated from the nature of the

proposed justification; the worth of agents must be derived solely from what it is that they can offer other agents – agents do not care about bargains reached on any other basis. And so we turned to an amalgamative justification, which has no such problem.

Amalgamative contractarianism is computationally trivial, but nevertheless cannot yet produce a justification because the necessary inputs are lacking. We do not have access to each agent's preferences. To proceed with a partial list of preferences, which we may more easily acquire, we would have to identify preferences that are of paramount import, and we would have to identify such preferences for each agent to whom morality is to apply. Such preferences are most reliably acquired through an investigation of contemporary psychological and economic results. Without a thorough examination of these fields, nothing can be said regarding the chances of success of such an enterprise. Without this knowledge, no justification can proceed, but we at least now have a focus for such an investigation. We know what we have to discover for any investigation of these fields to be considered a success.

Bibliography

Arrow, Kenneth, "Values and Collective Decision Making", *Rationality in Action*, P.K. Moser, ed., New York; Cambridge University Press, pp. 337-353.

Axelrod, Robert, *The Evolution of Cooperation*, New York; Basic Books, 1984.

Byron, Michael, "Satisficing and Optimality", *Ethics*, 109 (October 1998), pp. 67-93.

Campbell, Richmond, *Self Love & Self Respect: A Philosophical Study of Egoism*, Ottawa; Canadian Library of Philosophy, 1979.

Coleman, Jules & Morris, Christopher, *Rational Commitment and Social Justice: Essays for Gregory Kavka*, New York; Cambridge University Press, 1998.

Danielson, Peter, *Artificial Morality: Virtuous Robots for Virtual Games*, New York; Routledge, 1992.

Danielson, Peter (ed.), *Modeling Rationality, Morality, and Evolution*, New York; Oxford University Press, 1998.

DeJassay, Anthony, *Social Contract, Free Ride: A study of the public Goods Problem*, New York; Oxford University Press, 1990.

Dimock, Susan, "Defending Non-Tuism", *Canadian Journal of Philosophy*, 29:2 June 1999, pp. 251-274.

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

Fehige, Christoph & Wessels, Ulla (eds.), *Preferences*, New York; Walter de Gruyter & Co., 1998.

Frank, Robert H., *Choosing the Right Pond*, New York; Oxford University Press, 1985.

Friedman, Jeffrey (ed.), *The Rational Choice Controversy*, New Haven; Yale University Press, 1996.

Gardenfors, P. & Sahlin, N., *Decision, Probability, and Utility*, Cambridge University Press; Cambridge, 1988.

Gauthier, David & Sugden, Robert, *Rationality, Justice and the Social Contract: Themes From Morals by Agreement*, Ann Arbor; University of Michigan Press, 1993.

Gauthier, David, *The Logic of Leviathan*, Oxford at the Clarendon Press, 1969.

Gauthier, David, *Practical Reasoning*, Glasgow; Oxford University Press, 1963.

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

Gauthier, David, *Moral Dealing: Contract, Ethics, and Reason*, Ithica; Cornell University Press, 1990.

Gert, Bernard, *Morality: A New Justification of The Moral Rules*, New York; Oxford University Press, 1988.

Gert, Bernard, "Hobbes, Mechanism & Egoism", *Philosophical Quarterly*, 15(1965), pp. 341-349.

Gert, Bernard, "Hobbes and Psychological Egoism", *Journal of the History of Ideas*, 28(1967), pp. 503-520.

Grice, Russell, *The Grounds of Moral Judgment*, New York; Cambridge, 1967.

Hacking, Ian, *The Emergence of Probability*, Cambridge; Cambridge University Press, 1975.

Habermas, Jurgen, *Moral Consciousness and Communicative Action*, Cambridge; MIT Press, 1993.

Habermas, Jurgen, *Justification and Application: Remarks on Discourse Ethics*, Cambridge; MIT Press, 1993.

Habermas, Jurgen, *The Theory of Communicative Action* Vol. 1-2, Boston; Beacon Press, 1983.

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

Hobbes, Thomas, *Leviathan*, New York; Penguin Books, 1985.

Hobbes, Thomas, *Man and Citizen*, Bernard Gert (ed.), Indianapolis; Hackett Publishing Company, 1993.

Hobbes, Thomas, *The Correspondence of Thomas Hobbes Volume I*, Malcolm, Noel (ed), Oxford; Clarendon Press, 1997.

Hobbes, Thomas, *The Correspondence of Thomas Hobbes Volume II*, Malcolm, Noel (ed), Oxford; Clarendon Press, 1997.

Hurka, Thomas, "Consequentialism and Content", *American Philosophical Quarterly*, 29:1, January 1992, pp. 71-78.

Jeffrey, Richard C., *The Logic of Decision*, New York; McGraw-Hill Inc., 1965.

Johnston, David, *The Rhetoric of Leviathan*, New Jersey; Princeton University Press, 1986.

Kavka, Gregory, *Hobbesian Moral and Political Theory*, New Jersey; Princeton University Press, 1986.

Kraus, Jody, *The Limits of Hobbesian Contractarianism*, Cambridge; Cambridge University Press, 1993.

Lewis, Harry R., & Papadimitriou, Christos H., *Elements of the Theory of Computation*, New Jersey; Prentice-Hall Inc., 1981.

Luce, R. Duncan & Raiffa, Howard, *Games and Decisions: Introduction and Critical Survey*, New York; John Wiley & Sons, Inc., 1958.

Machover, Moshé, *Set theory, logic and their limitations*, Cambridge; Cambridge University Press, 1996.

MacIntosh, Duncan, "Two Gauthier's?", *Dialogue*, 1991, p. 3-32.

MacIntosh, Duncan, "Preference-Revision and the Paradoxes of Instrumental Rationality" *Canadian Journal of Philosophy* (22:4), Dec. 1992, pp. 503-530.

MacIntosh, Duncan, "Co-operative Solutions to the Prisoner's Dilemma", *Philosophical Studies*, (64:3) December 1991, pp. 309-321.

MacIntosh, Duncan, "Retaliation Rationalized: Gauthier's Solution to the Deterrence Dilemma", *Pacific Philosophical Quarterly*, (72:1) March 1991, pp. 9-32.

McNeilly, F. S., *The Anatomy of Leviathan*, Toronto; Macmillan, 1968.

Morris, Christopher, "The Relation Between Self-Interest and Justice in Contractarian Ethics" *The New Social Contract*, Paul, E., Miller, F, Paul, J., and Ahrens, J. eds. Basil Blackwell, 1988, pp. 119-153.

Nielsen, Kai & Shiner, Roger A., *New Essays on Contract Theory, Canadian Journal of Philosophy* Supplementary Volume 3, 1977.

Narveson, Jan, *The Libertarian Idea*, Philadelphia; Temple University Press, 1988.

Nozick, Robert, *Anarchy, State, and Utopia*, New York; Basic Books, Inc., 1974.

Paul, Ellen Frankel, Miller, Fred D. & Paul, Jeffrey (eds.), *Self-Interest*, Cambridge; Cambridge University Press, 1997.

Paul, Ellen Frankel, Miller, Fred D., Paul, Jeffrey & Ahrens, John (eds.), *The New Social Contract: Essays on Gauthier*, Basil Blackwell Ltd.: New York, 1988.

Rawls, John, *Collected Papers*, Freeman, S. (ed.), Cambridge, Massachusetts; Harvard University Press, 1999.

Rawls, John, *The Law of Peoples*, Cambridge, Massachusetts; Harvard University Press, 1999.

Rawls, John, *A Theory of Justice*, Cambridge, Massachusetts; Harvard University Press, 1971.

Rawls, John, *Political Liberalism*, New York; Columbia University Press, 1996.

Russell, Suart & Subramanian, Devika, "Provably Bounded-Optimal Agents", *Journal of Artificial Intelligence Research*, 2 (1995) pp. 575-609

Schmidtz, David, *Rational Choice and Moral Agency*, New Jersey; Princeton University Press, 1995.

Shaver, Robert, *Rational Egoism: A selective and critical history*, Cambridge; Cambridge University Press, 1999.

Simon, Herbert A., *Models of Man*, New York, John Wiley & Sons Inc., 1957.

Simon, Herbert A., *Administrative Behavior*, New York; The Macmillan Company, 1945.

Simon, Herbert A., *Models of Bounded Rationality*, Vol. 1-2, Cambridge; MIT Press, 1982.

Skyrms, Brian, *Evolution of the Social Contract*, Cambridge; Cambridge University Press, 1996.

Slote, Michael, *Beyond Optimizing: A Study of Rational Choice*, Cambridge, Massachusetts; Harvard University Press, 1989.

Sorell, Tom (ed.), *The Cambridge Companion to Hobbes*, New York; Cambridge University Press, 1996.

Taylor, Michael, *The Possibility of Cooperation*, Cambridge; Cambridge University Press, 1987.

Thagard, Paul, "Computational Tractability and Conceptual Coherence: Why do Computer Scientists Believe that P ≠ NP?", *Canadian Journal of Philosophy*, 23:3, Sept. 1993, pp. 349-364.

Thagard, Paul & Verbeurgt, Karsten, "Coherence as Constraint Satisfaction", *Cognitive Science*, 22:1 1998, pp. 1-24.

Vallentyne, Peter (ed.), *Contractarianism and Rational Choice*, New York; Cambridge University press, 1991.

Von Neumann, J. & Morgenstern, *Theory of Games and Economic Behavior* (3rd edition), Princeton, New Jersey; Princeton University Press, 1953.

Wicksteed, Philip Henry, *The Common Sense of Political Economy* (2 vols.), Lionel Robbins, (ed.), New York; Augustus M. Kelly (Publisher), 1967.

Winch, David M., *Analytical Welfare Economics*, Harmondsworth, England; Penguin Books Ltd., 1973.